Strand Passage and Knotting Probabilities in an Interacting Self-Avoiding Polygon Model

A Thesis Submitted to the
College of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Department of Mathematics and Statistics
University of Saskatchewan
Saskatoon

By
Matthew Frank Schmirler

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Mathematics and Statistics
Room 142 McLean Hall
106 Wiggins Road
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5E6

Abstract

The work presented in this thesis develops a new model for local strand passage in a ring polymer in a dilute salt solution. This model, called the Interacting Local Strand Passage (ILSP) model, models ring polymers via Θ-SAPs, which are self-avoiding polygons (SAPs) in the simple cubic lattice that contain a fixed structure Θ. This fixed structure represents two segments of the self-avoiding polygon being brought "close" together for the purpose of performing a strand passage. Θ-SAPs were first studied in the Local Strand Passage (LSP) model developed by Szafron (2000, 2009), where each Θ-SAP is considered equally likely in order to model good solvent conditions. In the ILSP model, each Θ-SAP has a modified Yukawa potential energy which contains an attractive term as well as a screened Coulomb potential that accounts for the effect of salt in the model. The energy function used in this model was first proposed by Tesi *et al.* (1994) for studying self-avoiding polygons in the simple cubic lattice.

The ILSP model is studied in this thesis using the Interacting Θ-BFACF (I-Θ-BFACF) Algorithm, an algorithm which is developed in this thesis and is proven to be ergodic on the set of all Θ-SAPs of a particular knot type and connection class. The I-Θ-BFACF algorithm was created by modifying the Θ-BFACF algorithm developed by Szafron (2000, 2009) to include energy-based Metropolis sampling. This modification allows one to sample Θ-SAPs of a particular knot type and connection class based on *a priori* chosen solvent conditions.

Multiple simulations (each consisting of 40 billion time steps) of composite Markov Chain Monte Carlo implementations of the I-Θ-BFACF algorithm are performed on unknotted connection class II Θ-SAPs (a.k.a Θ^{-}-SAPs) over a wide range of salt concentrations. The data from these simulations is used to estimate, as a function of polygon length, the probability of an unknotted Θ^{-}-SAP remaining an unknot after a strand passage, as well as the probability of it becoming a positive trefoil knot. The results strongly suggest that as the length of a Θ-SAP goes to infinity, the probability of the Θ-SAP becoming knotted after a strand passage increases as the salt concentration in the model increases. These results serve as a first step for studying how the knot reduction factor (studied by Liu *et al.* (2006) and Szafron and Soteros (2011)) of a ring polymer varies in differing solvent conditions. The goal of this future research is to find solvent conditions and a local geometry of the strand passage site that yields a knot reduction factor comparable to the research of Rybenkov *et al.* (1997), which shows an 80-fold reduction of knotting after type II topoisomerase enzymes act on DNA.

Acknowledgements

I would like to thank my supervisor, Dr. Chris Soteros, for her expertise, advice, and the many hours that were spent reading the various drafts of this thesis. I am also very grateful for the financial support and opportunities provided to me by her NSERC Discovery grant. I would also like to thank Dr. Michael Szafron for letting me use his Θ-BFACF algorithm code and for his invaluable advice on many complicated issues pertaining to this thesis. I greatly appreciate the input and suggested improvements to my thesis given by my advisory committee members: Dr. Szafron, Dr. Soteros, as well as Dr. M. Bickis, Dr. Longhai Li, and Dr. T. Kusalik. I would also like to thank Dr. Mariel Vazquez, Dr. S.G. Whittington, Dr. E. Orlandini and Dr. C. Tesi for their advice on relevant issues pertaining to this thesis. I am endebted to Dr. Rob Scharein for his friendship and for allowing me to use *KnotPlot* for its HOMFLY polynomial, as well as for its amazing graphics capabilities which were used to create many of the images used in this thesis. I would also like to thank Compute Canada for allowing me to use their high performance computing network, and the College of Graduate Studies and Research for their support in the form of a Dean's Scholarship. I would like to thank Marla Cheston and Kevin McGregor for providing me with a working knot identification program, as well as for their friendship and advice. I would also like to thank my family, friends, and the Department of Mathematics and Statistics staff for all their support; without this support it is unlikely that I would have finished my thesis (in a timely manner).

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The work presented in this thesis is motivated by two main problems:

Problem 1. *What is a 'good way' to model ring polymers that exist in a salt solution?*

Problem 2. *How do the knotting probabilities relating to strand passages within an unknotted ring polymer change with the concentration of salt in the solution?*

Problem 1 is motivated by experimental evidence obtained by Shaw and Wang in [53] and by Rybenkov *et al.* in [50] which shows that the probability of a cyclized DNA molecule in solution being knotted increases significantly with the salt concentration of the solution. A main goal of this work is to simulate a model of ring polymers in salt solution that will provide results whose trends qualitatively reproduce those obtained in the experiments of [50] and [53].

Up to date, there have been several random polygon models [6, 17, 27, 36, 37, 38, 60, 61, 62, 63, 64] designed to simulate DNA in solution (this is by no means an exhaustive list); these polygon models range from worm-like chain models, considered more DNA-like, to lattice models, which provide a much coarser approximation. Several of these models [6, 17, 27, 37, 60, 61] operate under the assumption that the ring polymers are in solution with a *good solvent* (*i.e.* the salt concentration is negligible). However, under physiological conditions, DNA exists in a salt solution [54]. DNA is also negatively charged and interacts with counterions that exist in such a salt solution [3]; therefore, it would be desirable to model these interactions somehow. Only a few of the models listed above [36, 38, 64] contain some type of energetic term that tries to take into account the interactions that occur with polymers in salt solution. Notably, in [64] they use a lattice model, and using a Markov Chain Monte Carlo (MCMC) algorithm known as the pivot algorithm they are able to obtain knot probabilities whose trends closely relate to those of the experiments of [53].

Problem 2 stems from Problem 1, and is motivated by the desire to better understand the function of type II topoisomerase enzymes on DNA. Type II topoisomerase enzymes have the ability to pass one segment of double stranded DNA through another by cleaving and opening one DNA

duplex, passing a second duplex through the opening, and re-ligating the break [67]. This *segment passage* action is performed at a local site on the DNA, but has the ability to change the knot type of the molecule, which is a global property. Experimental results show that type II topoisomerase enzymes have the ability to reduce knotting of DNA molecules at steady state to as much as 80 times lower than what it would be at thermodynamic equilibrium [51]. Moreover, the action of these enzymes is vital to a cell: the absence of type II topoisomerase enzymes at mitosis or meiosis will ultimately cause cell death [69]. Some anticancer drugs used in chemotherapy try to exploit this by inhibiting the action of type II topoisomerases [74].

Although it is known that type II topoisomerase enzymes perform strand passages, exactly how the enzyme chooses where to act on the DNA molecule is not. Understanding how type II topoisomerases choose where to act on DNA, and how they unknot DNA so successfully is still an open question in molecular biology [62]. There has been progress made on these questions; the work of Mann, [42] Mann *et al.* [37, 43], Neuman *et al.* [44], and Szafron and Soteros [62] suggests that this enzyme is not acting at a random location on DNA, rather that it is acting preferrentially depending on local geometry at the strand passage site.

In an attempt to model type II topoisomerase induced strand passages in random ring polymers, Szafron [60, 61] devised a model (called the *Local Strand Passage Model* [61]) in which all modelled polymers contain two fixed segments that are 'pinched' together for the purpose of performing a strand passage. In [61], Szafron estimates knotting probabilities relating to a single strand passage at a specified strand passage structure within an unknotted ring polymer that is assumed to be in a good solvent. Szafron and Soteros [62, 63] have also explored how these strand passage probabilities are affected by the local geometry of the fixed segments in the model. The work in [60, 61, 62, 63] used a MCMC method known as a Composite Markov Chain (CMC) implementation of the Θ-BFACF algorithm to study this model. In order to address the strand passage probabilities mentioned in Problem 2, the work presented here extends the Local Strand Passage (LSP) Model to include the interactive model for polymers in salt solution used in [64]. This new model is called the interacting LSP model, or ILSP model for short. To study this new model, the Θ-BFACF algorithm is modified to include Metropolis sampling based on solvent conditions.

The first chapter of this thesis will review some of the basic terminology relating to ring polymers, how they are modelled here (via *self-avoiding polygons* in the simple cubic lattice), as well as what it means mathematically for a ring polymer to be 'knotted'. Chapter 1 will also further discuss the action of the the type II topoisomerase enzyme on DNA and the consequences of this action

on the knot type of DNA. Chapter 2 is devoted to explaining how the action of this enzyme is incorporated into the Local Strand Passage Model [61], theoretical results relating to this model, as well as definitions for some observable quantities of interest relating to the model. Chapter 3, and a majority of Chapter 4, are devoted to reviewing the statistical theory and algorithms that can be used to generate 'essentially independent' samples of random self-avoiding polygons. The purpose of this review is to present some of the major theoretical results which are crucial in establishing the ILSP model. A major assumption of the algorithms being reviewed is that the polygons are in a dilute solution with a '*good solvent*'. In Chapter 5, the method of Metropolis sampling is reviewed. Using this method, it is described how to modify the pivot algorithm and the Θ-BFACF algorithm in order to generate a sample of essentially independent self-avoiding polygons from distributions relating to particular solvent conditions. These modified algorithms are referred to here as the *Interacting Pivot Algorithm* (also referred to here as the *I-Pivot Algorithm*) and the *Interacting* Θ*-BFACF Algorithm* (also referred to as the *I-*Θ*-BFACF Algorithm*), respectively. It should be clarified that the I-Pivot Algorithm used here was first presented by Tesi *et al.* in [64]. Chapter 6 reviews the techniques that are necessary to analyze data coming from the algorithms described in Chapters 4 and 5, as well as some specific methods for computing strand passage and knot transition probabilities in the LSP and ILSP models. These techniques are essential to answering Problems 1 and 2. Chapter 7 assesses the consistency of the algorithms used here. By considering 'good solvent' conditions in the ILSP model, a CMC implementation of the new I-Θ-BFACF Algorithm is directly compared to Szafron's CMC implementation of the Θ-BFACF algorithm in [61]. Also, as I independently programmed the I-Pivot Algorithm, some of the results obtained from simulations of this algorithm are directly compared with the results obtained in [64]. In order to address Problem 1, the results of the I-Pivot Algorithm are also compared to the experimental results obtained by Shaw and Wang in [53]. Chapter 8 addresses Problem 2 by presenting some knot transition probability results for different solvent conditions in the ILSP model via CMC implementations of the I-Θ-BFACF Algorithm. Appendix C contains a list of definitions for symbols that are used throughout the thesis.

1.1 Definitions

The following section will introduce some basic definitions necessary to understand how ring polymers are modelled in this work.

The definitions in this paragraph are based on those given in [40]. A *polymer* is a molecule that consists of many repeated *monomers* (groups of atoms) joined together by chemical bonds. The *functionality* of a monomer is the number of available sites for chemical bonds that it has, that is, the maximum number of other monomers with which it can bond. A *linear polymer* is a chain of monomers with functionality two, terminated at both ends by monomers with functionality one. A *ring polymer* is a chain of monomers where each monomer has functionality two with the monomers on the ends bonded to each other. If a particular polymer is made up of two or more different types of monomers, it is referred to as a *copolymer*; if all the monomers are the same, it is referred to as a *homopolymer*. A single strand of deoxyribonucleic acid (DNA) in humans is an example of a copolymer comprised of four different types of monomers [3] (*i.e.* nucleotides).

"In 1963, Dulbecco and Vogt [13] and Weil and Vinograd [73] discovered that double-stranded DNA of the polynoma virus exists in a closed circular form. At present, it is generally acknowleged that this form is typical of bacterial DNA and of cytoplasmic DNA in animals" [72, page 1]. At a macroscopic scale, circular double-stranded DNA can be thought of simply as a *ring homopolymer* if one focusses on the axis that the DNA double helix winds around. This is the viewpoint taken in this thesis; the reason for this is that we are interested in studying knotting, which is defined for a simple closed object.

Because the DNA of animals and humans is linear, it would seem that the models presented here are not relevant to animals or humans. This is not necessarily true, because "giant DNA molecules in higher organisms form loop structures held together by protein fasteners in which each loop is largely analogous to closed circular DNA" [72, page 1]. Furthermore, "the distinctive feature of closed circular molecules is that their topological state cannot be altered by any conformational rearrangement short of breaking the DNA strands" [72, page 1]. What this tells us is that in essence both linear and circular DNA have topological conformations that are subject to possible constraints, one of which is that the DNA can become knotted.

In the work presented here, ring polymers are modelled by *self-avoiding polygons* (also referred to here as *SAPs* or simply *polygons*) in the simple cubic lattice (to be defined in Section 1.3). In the LSP and ILSP models, the polygons considered will have a particular fixed structure (called the Θ-structure, described in Section 2.1.1) which represents two segments of the polygon being brought close together. In these models, the Θ-structure in a polygon can be replaced with an alternate structure. This replacement procedure has the effect of performing a strand passage on the polygon and models the effect of the type II topoisomerase action.

In summary, a ring polymer is a closed circuit of monomers. Circular DNA can be thought of as a ring polymer, while linear DNA can form loop structures that are essentially equivalent to ring polymers. In the work presented here, ring polymers are modelled by self-avoiding polygons in the simple cubic lattice, and the action of the type II topoisomerase enzyme is modelled by replacing a fixed structure in a polygon with another structure. This replacement procedure can change the knot type of a polygon, but what exactly does this mean? Also, what does it mean for two knots to have different knot types? These questions will be addressed and answered in the following section.

1.2 Knot Theory - What is a Knot?

When most people think of a knot, they might think of a shoelace, a tie, or a rope. To define a knot mathematically, we need to be more careful. Because any shoelace, tie, or rope knot can be undone, there is a need to introduce restrictions that help identify when one type of knot is different from another. If a knot is forced to be a closed, non self-intersecting curve in $\mathbb{R}^3$, we can define two knots to be different if one cannot be smoothly deformed into the other [49].

Knot theory comes up in many different and surprising areas, both naturally and man-made. The following is a very brief introduction into this broad area of research.

1.2.1 Standard Terminology

The following section contains standard definitions and terminology used in knot theory; unless otherwise stated, the following definitions come from [49].

A mapping $f : \mathbb{S}^1 \to \mathbb{R}^3$ is called an *embedding of the unit circle into* $\mathbb{R}^3$. Such an embedding is referred to as a *knot* if it is homeomorphic to the unit circle [10]. Two knots K_1 and K_2 are defined to be *equivalent* if one can be continuously deformed into the other. The set of all knots that are equivalent to each other form *equivalence classes* known as *knot types*. Figure 1.1 shows examples of three of the simplest knot types in their simplest form. From left to right these knot types are: the (trivial) unknot (denoted by ϕ), the trefoil knot (denoted by 3_1) and the figure-8 knot (denoted by 4_1) [10]. A knot is said to be *chiral* if it is not equivalent to its mirror image. If a knot is not chiral, it is said to be *achiral*. The trefoil is an example of a chiral knot, whereas the figure-8 is an example of an achiral knot. For the purpose of distinguishing between the two chiralities of trefoils, define 3_1^+ to be the knot type of the left image in Figure 1.2, and define 3_1^- to be the knot type of the right image in Figure 1.2.

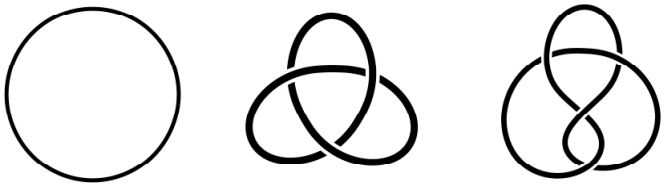

Figure 1.1: From left to right, knot diagrams of the unknot (ϕ), trefoil (3_1) and figure-8 (4_1) knots. Images created using *KnotPlot* [52].

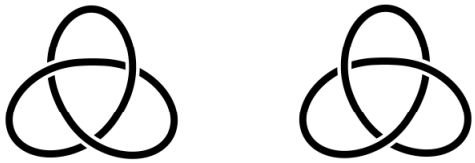

Figure 1.2: The trefoil is a chiral knot (*i.e.* it is not equivalent to its mirror image). The knot on the left is an example of a 3_1^+ knot, while the knot on the right is a 3_1^- knot. Images created using *KnotPlot* [52].

A *polygonal knot* is a knot in $\mathbb{R}^3$ which is made up of a finite collection of linear segments called *edges*. The endpoints of these edges are called *vertices*.

One important issue is how to determine the knot type of a knot. For this purpose, it is often useful to convert a knot in $\mathbb{R}^3$ to a two dimensional projection. The following terminology and theory is based on [5]: A projection $\varphi : \mathbb{R}^3 \to \mathbb{R}^2$ of a knot K is said to have a *multiple point* at $a \in \mathbb{R}^2$ if $\varphi^{-1}(a) \cap K$ consists of more than one point. A multiple point where $\varphi^{-1}(a) \cap K$ consists of two points is defined to be a *double point*. A projection φ of a knot K is considered to be a *regular projection* if $\varphi(K)$ has a finite number of multiple points, where each multiple point is a double point where two segments of K cross transversely (see Figure 1.3 for examples of proper and improper crossings in a projection). If K is a polygonal knot, the requirements for φ being a regular projection are the same, with the added condition that no double point in $\varphi(K)$ is the image of a vertex of K.

A knot K is said to be in *regular position* if there exists a regular projection of K onto the xy-plane. Is it always possible to transform a polygonal knot into a polygonal knot in regular position? The following theorem and its corollary indicate that the answer is "yes".

Theorem 1.2.1 ([47])**.** *If K is a polygonal knot, then there is an arbitrarily small rotation of $\mathbb{R}^3$ onto $\mathbb{R}^3$ that maps K into a polygonal knot in regular position.*

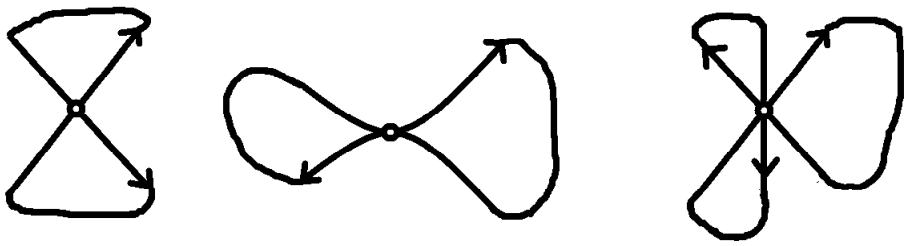

Figure 1.3: For a projection to be regular, all crossings in the projection must represent two segments crossing transversely. The left-most projection is an example of a regular projection. The center projection is not a regular projection because it contains a crossing that is not transverse. The right-most projection is not a regular projection because it has three strands crossing a single point.

Theorem 1.2.1 provides the following corollary:

Corollary 1.2.2 ([47])**.** *Every polygonal knot is equivalent to a polygonal knot in regular position.*

Suppose now that K is a knot in regular position. Without loss of generality, assume that $\varphi : \mathbb{R}^3 \to \mathbb{R}^2$ is defined by $\varphi((x, y, z)) = (x, y)$. By definition, each double point of $\varphi(K)$ is the mapping of exactly two points of K. The location of these double points are defined to be *crossings*. Whichever of these points has the larger z-coordinate is defined to be the *overcrossing*; define the other point to be the *undercrossing*. The segments of the projection of K which run through these points are defined to be the *overcrossing* and *undercrossing segments*, respectively.

Define a *knot projection* [5] to be a regular projection where the segments at every crossing are specified as overcrossings or undercrossings respectively. The image of this knot projection is referred to as a *knot diagram*. Define the *crossing number* of a knot to be the minimum number of crossings over all knot projections of the knot.

1.2.2 How Do We Tell When Two Knots Are Equivalent?

It is not always a simple task to tell when two knots are equivalent. For example, the knot pictured in Figure 1.4 is actually equivalent to the unknot, that is, it can be continuously deformed into a circle in $\mathbb{R}^3$. Is there a "nice" method to tell when two knots are the same? The remainder of this section will address this question.

Two knot projections φ_1 and φ_2 are called *equivalent projections* if there exists a finite sequence of *Reidemeister moves* [48] which transforms φ_1 into φ_2 (and vice versa). There are three types of Reidemeister moves, all of which are designed to change the positioning of crossings in a knot diagram. The first Reidemeister move (denoted Ω_1) allows a twist to be added or removed from a

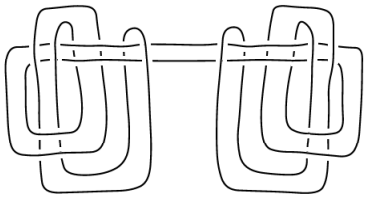

Figure 1.4: The 'nasty knot'; image created using the software *KnotPlot* [52].

knot diagram (see Figure 1.5). The second Reidemeister move (denoted Ω_2) allows for one segment in the knot diagram to be moved over/under another, and vice versa (see Figure 1.6). The final Reidemeister move (denoted Ω_3) allows for a segment in the diagram to slide from one side of a crossing to the other (see Figure 1.7). These three moves allow us to determine the following result.

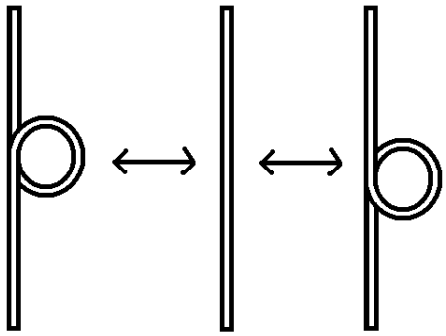

Figure 1.5: Reidmeister move Ω_1

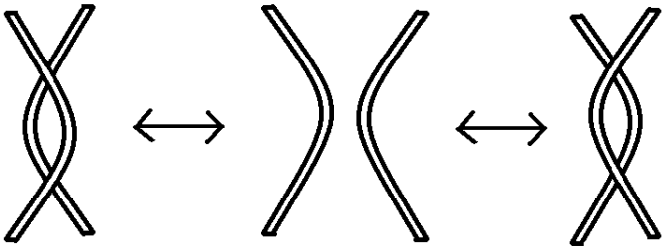

Figure 1.6: Reidmeister move Ω_2

Theorem 1.2.3 ([49])**.** *Two knots K_1 and K_2 have the same knot type if and only if a knot projection of K_1 is equivalent to a knot projection of K_2, that is, if and only if there is a finite sequence of Reidemeister moves that transforms the knot projection of K_1 into the knot projection of K_2.*

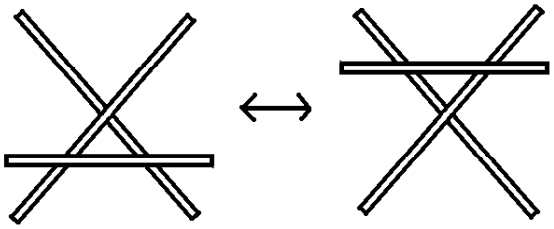

Figure 1.7: Reidmeister move Ω_3

A consequence of Theorem 1.2.3 is that if two knots have the same knot type, one can find a sequence of Reidemeister moves to convert the knot diagram of one knot into the knot diagram of the other. However, finding such a sequence can be very difficult. In practice, one uses instead a computational algorithm such as the Alexander Polynomial [1], the Jones Polynomial [29], or the HOMFLY polynomial [15] in order to attempt to identify a knot diagram's knot type (in actuality, the Jones and Alexander Polynomials are both special cases of the HOMFLY polynomial [2]). In this work, the HOMFLY polynomial implemented in *KnotPlot* [52] is used. The HOMFLY polynomial is a 2 variable knot polynomial that is knot invariant, that is, all knots with the same knot type will have the same HOMFLY polynomial.

The HOMFLY polynomial does have limitations. One limitation is that the HOMFLY polynomial is not always able to distinguish between different knot types: for example, the knot types 5_1 and 10_{132} yield the same HOMFLY polynomial [30]. Another limitation of the HOMFLY polynomial is that it cannot always distinguish between the *chirality* (*i.e.* handedness) of the knot: for example, the 9_{42} knot type and its mirror image have the same HOMFLY polynomial [58]. Because in the work presented here knot types with five or more crossings are rarely observed, the error due to these HOMFLY polynomial limitations is negligible (see Section 8.8 for justification).

1.3 Lattice Theory - How do we Model Polymers?

As mentioned in the introduction of this chapter, one can model double stranded circular DNA using self-avoiding polygons in the simple cubic lattice, where the polygon represents the axis that the double helix of the DNA winds around. Although this is a rough model for DNA, one advantage is that this model can easily accommodate for the excluded volume property of DNA. Another advantage of this lattice model is that it is amenable to combinatorial and asymptotic analysis [62]. The following section will precisely define what is meant by a self-avoiding polygon in the simple

cubic lattice, as well as some of the major theoretical results pertaining to this model.

1.3.1 Basic Definitions

The following definitions are based on [40], unless otherwise stated.

Definition 1.3.1. *Define the* d-dimensional hypercubic lattice $\mathbb{Z}^d$ *to be the set of points*

$$V(\mathbb{Z}^d) := \{(x_1, \ldots, x_d) | x_1, \ldots, x_d \in \mathbb{Z}\} \tag{1.1}$$

as well as the set of edges

$$E(\mathbb{Z}^d) := \left\{ \{\mathbf{x}, \mathbf{y}\} | \mathbf{x}, \mathbf{y} \in V(\mathbb{Z}^d), \sum_{i=1}^{d} |x_i - y_i| = 1 \right\}. \tag{1.2}$$

The work presented in this thesis uses the 3-dimensional hypercubic lattice, denoted by $\mathbb{Z}^3$, which is also known as the *simple cubic lattice*.

Definition 1.3.2. *An* n-edge self-avoiding walk (SAW) u *in* $\mathbb{Z}^3$ *beginning at site* $\boldsymbol{x}$ *is defined to be a directed graph embedding* $u = (V(u), E(u))$ *in* $\mathbb{Z}^3$ *consisting of a sequence of* n *distinct arcs in* $\mathbb{Z}^3$, $E(u) = ((\boldsymbol{v}_1, \boldsymbol{v}_2), (\boldsymbol{v}_2, \boldsymbol{v}_3), \ldots, (\boldsymbol{v}_n, \boldsymbol{v}_{n+1}))$, *and a corresponding sequence of* $n+1$ *distinct vertices in* $\mathbb{Z}^3$, $V(u) = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{n+1})$, *such that the vertices* $\boldsymbol{v}_i \in \mathbb{Z}^3$ *for each* $i = 1, \ldots, n+1$, $\boldsymbol{v}_1 = \boldsymbol{x}$, *and for each* $i = 1, \ldots, n$, *the arc* $(\boldsymbol{v}_i, \boldsymbol{v}_{i+1})$ *joins two adjacent vertices in* $\mathbb{Z}^3$. *The* length *of this self-avoiding walk is defined to be the number of arcs in* $E(u)$.

Definition 1.3.3 ([61])**.** *A* $2n$-edge self-avoiding polygon (SAP) ω, *for* $n \geq 2$ *is defined to be a graph embedding* $\omega = (V(\omega), E(\omega))$ *in* $\mathbb{Z}^3$ *consisting of a set of* $2n$ *distinct edges in* $E(\mathbb{Z}^3)$,

$$E(\omega) = \{\{\mathbf{v}_0, \mathbf{v}_1\}, \{\mathbf{v}_1, \mathbf{v}_2\}, \ldots, \{\mathbf{v}_{2n-2}, \mathbf{v}_{2n-1}\}, \{\mathbf{v}_{2n-1}, \mathbf{v}_0\}\}, \tag{1.3}$$

and a corresponding set of $2n$ *distinct vertices in* $\mathbb{Z}^3$,

$$V(\omega) = \{\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_{2n-1}\}. \tag{1.4}$$

Define the length *of* ω *to be the number of edges in* $E(\omega)$.

From this definition it is clear that a SAP is a type of polygonal knot, specifically one whose edges and vertices are strictly in $\mathbb{Z}^3$.

Definition 1.3.4. *For any SAP or SAW* ω, *define* $|\omega|$ *to be the length of* ω.

Definition 1.3.5. *For any SAP ω, define $k(\omega)$ to be the knot type of ω.*

Definition 1.3.6. *Define $\mathcal{C}_n$ to be the set of all n-edge SAWs in $\mathbb{Z}^3$ with $\boldsymbol{v}_1 = (0,0,0)$. Let $c_n = |\mathcal{C}_n|$ be the number of n-edge SAWs in $\mathbb{Z}^3$ with $\boldsymbol{v}_1 = (0,0,0)$. Define $\mathcal{C} = \bigcup_{n=1}^{\infty} \mathcal{C}_n$ to be the set of all SAWs in $\mathbb{Z}^3$ with $\boldsymbol{v}_1 = (0,0,0)$.*

Definition 1.3.7 ([61])**.** *For a SAP $\omega \in \mathbb{Z}^3$, ω is referred to as a* rooted polygon *if one of its vertices is designated as the* root *of ω. If no vertex of ω is specified as the root, then ω is referred to as an* unrooted polygon.

Definition 1.3.8. *Define $\mathcal{P}_{2n}$ to be the set of all $2n$-edge SAPs in $\mathbb{Z}^3$. Let p_{2n} be the number of $2n$-edge SAPs in $\mathbb{Z}^3$ up to translation. Define $\mathcal{P} = \bigcup_{n=1}^{\infty} \mathcal{P}_{2n}$ to be the set of all SAPs in $\mathbb{Z}^3$.*

Since we are sometimes interested in particular knot types, it is useful to partition $\mathcal{P}_n$ and $\mathcal{P}$ accordingly.

Definition 1.3.9. *For any knot type K, define $\mathcal{P}_{2n}(K) = \{\omega | \omega \in \mathcal{P}_{2n},\, k(\omega) = K\}$. Let $p_{2n}(K)$ be the number of $2n$-edge SAPs with knot type K (up to translation). Define $\mathcal{P}(K) = \bigcup_{n=1}^{\infty} \mathcal{P}_{2n}(K)$.*

1.3.2 Asymptotic Behaviour of $\mathbf{p_{2n}}$ and $\mathbf{c_n}$

Although we can precisely define what $\mathcal{P}_{2n}$ and $\mathcal{C}_n$ are, determining p_{2n} and c_n is computationally demanding. For example, Table 1.1 contains a list of values of n for which p_{2n} and c_n have been enumerated [9]. The rapid increase of p_{2n} and c_n shown in Table 1.1 indicates that computing p_{2n} and c_n as $n \to \infty$ is not feasible; in fact, as of 2007, the largest completely enumerated value of p_{2n} is for $2n = 32$ [9].

n	p_n	c_n
4	3	726
10	2,412	8,809,878
20	1,768,560,270	49,917,327,838,734
30	2,912,940,755,956,084	270,569,905,525,454,674,614

Table 1.1: A few values of n for which p_n and c_n have been enumerated [9].

So how does c_n grow with n? Hammersley and Morton [24] proved that c_n grows at an exponential rate:

Theorem 1.3.10 ([24]). *The following limit exists:*

$$0 < \kappa := \lim_{n\to\infty} \frac{1}{n} \log c_n < \infty, \tag{1.5}$$

where κ is referred to as the connective constant *for SAWs in $\mathbb{Z}^3$.*

Definition 1.3.11. *The quantity $\mu := e^{\kappa}$ is referred to as the* growth constant *for SAWs in $\mathbb{Z}^3$.*

Also of interest is determining the asympotic form of p_{2n}; Hammersley [23] proved that p_{2n} not only grows exponentially, but that it also grows at the same exponential rate as c_n.

Theorem 1.3.12 ([23]). *The following limit exists and satisfies:*

$$\lim_{n\to\infty} \frac{1}{n} \log p_{2n} = \kappa. \tag{1.6}$$

There are some results that have been proven for the connective constants and growth rates of $p_n(K)$ for different knot types. For the unknot, the following result was proved by Sumners and Whittington [59]:

Theorem 1.3.13 ([59]).

$$0 < \lim_{n\to\infty} \frac{1}{2n} \log p_{2n}(\phi) =: \kappa_\phi < \kappa. \tag{1.7}$$

Definition 1.3.14. *The quantity $\mu_\phi := e^{\kappa_\phi}$ is referred to as the* growth constant *for unknotted SAPs in $\mathbb{Z}^3$.*

Theorem 1.3.13 states that the exponential growth rate of unknotted SAPs is less than the exponential growth rate of all SAPs. This implies that as $n \to \infty$, the probability of a $2n$-edge SAP being knotted goes to 1.

Unfortunately, a limit for any non-trivial knot type K defined analogously to that in Equation 1.7 has not yet been proven to exist. Soteros, Sumners and Whittington [57] proved a weaker result for SAPs with a non-trivial knot type K:

Theorem 1.3.15 ([57]). *For any knot type K,*

$$k_K := \liminf_{n\to\infty} \frac{1}{2n} \log p_{2n}(K) \leq \limsup_{n\to\infty} \frac{1}{2n} \log p_{2n}(K) =: \kappa_K < \kappa = \log \mu, \tag{1.8}$$

with

$$\kappa_K \geq \kappa_\phi. \tag{1.9}$$

It is also believed [46] that $\kappa_K = \kappa_\phi$, but proving the inequality $\kappa_K \leq \kappa_\phi$ is still an open question.

Theorem 1.3.12 establishes that p_{2n} grows exponentially. It is believed that there exists constants A, α, μ, B, and Δ such that p_{2n} has the asymptotic scaling form given by [46]:

$$A(2n)^{\alpha-3}\mu^{2n}\left(1+\frac{B}{(2n)^{\Delta}}+O(n^{-1})\right), \tag{1.10}$$

where A is the *amplitude*, $\mu = e^{\kappa}$, α is the *entropic critical exponent*, and Δ is a scaling correction referred to as a *confluent exponent*.

Because Theorem 1.3.13 establishes that $p_{2n}(\phi)$ also grows exponentially, Orlandini *et al.* [46] proposed that there are constants A_ϕ, α_ϕ, μ_ϕ, B_ϕ, and Δ_ϕ such that $p_{2n}(\phi)$ scales like:

$$A_\phi(2n)^{\alpha_\phi-3}\mu_\phi^{2n}\left(1+\frac{B_\phi}{(2n)^{\Delta_\phi}}+O(n^{-1})\right), \tag{1.11}$$

where A_ϕ is the amplitude, $\mu_\phi = e^{\kappa_\phi}$, α_ϕ is the entropic critical exponent and Δ_ϕ is the confluent exponent.

Assuming that $k_K = \kappa_K$ (as defined in Equation 1.8) for a non-trivial knot type K, Orlandini *et al.* [46] also proposed that $p_{2n}(K)$ has an analogous scaling form; *i.e.*

$$A_K(2n)^{\alpha_K-3}\mu_K^{2n}\left(1+\frac{B_K}{(2n)^{\Delta_K}}+O(n^{-1})\right), \tag{1.12}$$

where A_K is the amplitude, $\mu_K = e^{\kappa_K}$, α_K is the entropic critical exponent and Δ_K is the confluent exponent. It is also conjectured in [46] that

$$\kappa_K = \kappa_\phi, \ \forall\ K, \tag{1.13}$$

and

$$\alpha_K = \alpha_\phi + n_K, \tag{1.14}$$

where n_K is a knot-dependent constant that is believed to be the number of factors in the prime knot decomposition of the knot type K.

1.4 Chapter Summary

In this chapter, some motivation has been presented for developing models of ring polymers, and it has been introduced how self-avoiding polygons in the simple cubic lattice can model ring polymers.

Although the nature of the simple cubic lattice leads to a coarse approximation of a ring polymer, "lattice models have the advantage that they can easily incorporate the excluded volume property and are amenable to combinatorial and asymptotic analysis" [62, page 2]. Some results relating to such combinatorial and asymptotic analyses were described in detail in Section 1.3; these results will play an important role in the theory of the LSP and ILSP models.

Section 1.2 provided a mathematical basis for discriminating between different types of knots that can occur in a ring polymer. In the introduction of the chapter, it was described how such knots can occur in both linear and circular DNA, and how these knots can be removed by the strand passage action of type II topoisomerase enzymes. In fact, the action of these enzymes is vital; the absence of type II topoisomerase enzymes at meiosis or mitosis ultimately causes cell death [69]. Although the result of the action of type II topoisomerases on DNA is known, understanding how they choose where to act on DNA, and how they unknot DNA so successfully are still open questions in molecular biology.

It was also described how several models have been proposed to study the action of type II topoisomerases, and how many of these models assume that the ring polymers exist in a good solvent, something that is not true in the physiological conditions of DNA. The next chapter will explain how the local strand passage action performed by these enzymes can be modelled in the simple cubic lattice via the LSP model, as well as introduce an energy associated with self-avoiding polygons that can take into account the effect of a salt solution in the model.

Chapter 2

Modelling Strand Passage in Self-Avoiding Polygons

In order to learn more about type II topoisomerase enzymes, we need to have a model that can mimic its strand passage action. The following chapter reviews one such model for this action, called the *Local Strand Passage* model, as well as some theoretical results and observable quantities of interest relating to this model. The LSP model forms the basis for the new ILSP model presented in this thesis; the interaction energy used in the ILSP model is also defined in Section 2.3.2.

2.1 The Local Strand Passage (LSP) Model

The Local Strand Passage (LSP) Model of [60] and [61] was designed to study the effects of a strand passage in a ring polymer. The ring polymers (i.e. SAPs) in this model contain a fixed structure which represents two strands of the polymer being pinched closely together for the purpose of performing a strand passage. This 'pinched together' portion is modelled in SAPs by way of a fixed structure which forces two strands to be close together. This fixed structure, used by Szafron in [60] and [61] and by Szafron and Soteros in [62, 63], is referred to as the Θ-structure. A strand passage can be attempted by replacing the Θ-structure with an alternate structure. This replacement procedure models the strand passage that occurs due to the action of the type II topoisomerase enzyme in that it is equivalent to breaking one strand apart, passing another strand through, and resealing the break.

2.1.1 Definition of the Fixed Structure

Definition 2.1.1. *Define the Θ-structure (shown in Figure 2.1) to be the vertices $V(\Theta)$ and edges $E(\Theta)$, where:*

$$V(\Theta) = \{(-1,0,0),(0,0,0),(1,0,0),(0,-1,-2),(0,-1,-3),(0,0,-3),(0,1,-3),(0,1,-2)\}\,,$$

and

$$E(\Theta) = \{\{(-1,0,0),(0,0,0)\},\{(0,0,0),(1,0,0)\},\{(0,-1,-2),(0,-1,-3)\},$$
$$\{(0,-1,-3),(0,0,-3)\},\{(0,0,-3),(0,1,-3)\},\{(0,1,-3),(0,1,-2)\}\}.$$

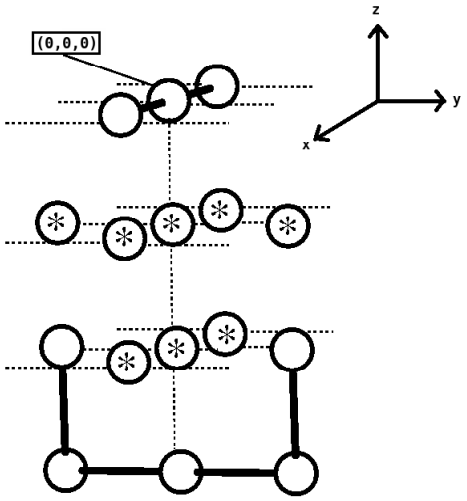

Figure 2.1: The Θ-structure in $\mathbb{Z}^3$. Empty circles represent the vertices of Θ, and solid lines represent the edges of Θ. Circles with an asterisk represent vertices which must be unoccupied for a strand passage to be successful.

Definition 2.1.2. *Define a Θ-SAP to be any self-avoiding polygon which contains the Θ-structure. An example of a Θ-SAP is shown in Figure 2.2.*

Any Θ-SAP can be broken down into the Θ-structure and two mutually-avoiding undirected self-avoiding walks connecting the segments of the Θ-structure together (see Figure 2.2). Θ-SAPs whose decomposition consists of the Θ-structure, a SAW connecting (-1,0,0) to (0,1,-2), and another SAW connecting (0,-1,-2) to (1,0,0) are defined to be elements of the set σ_1 [60], referred to here as *class I Θ-SAPs*. Any Θ-SAP which is not a class I Θ-SAP must have SAWs connecting (-1,0,0) to (0,-1,-2) and (0,1,-2) to (1,0,0). These Θ-SAPs are defined to be elements of the set σ_2 [60], referred to here as *class II Θ-SAPs*. The difference between these two classes also corresponds to a difference in the crossing sign of the two strands of the suitably oriented Θ-structure when projected onto the xy-plane; Figure 2.3 shows the convention by which crossing signs are assigned. For example, the projection of the Θ-structure as it occurs in the oriented Θ-SAP in Figure 2.2 corresponds to a positive crossing. For this reason, the two different classes of Θ-SAPs (I and II) are also referred to in [62] and [63] as being Θ^+- and Θ^--SAPs, respectively.

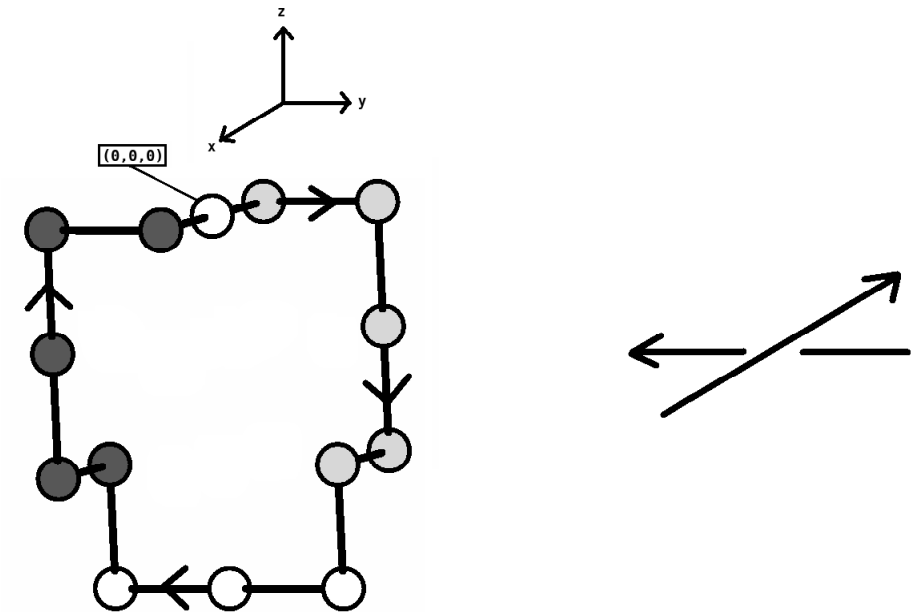

Figure 2.2: The left image is an example of a Θ-SAP (i.e. a SAP with the Θ-structure); the shaded vertices correspond to the two mutually-avoiding undirected SAWs connecting the two segments of the Θ-structure. The right image shows the projection of the top and bottom segments of the Θ-structure onto the xy-plane. Θ-SAPs in this class are referred to as Θ^{+}-SAPs because the projection of the Θ-structure yields a positive crossing.

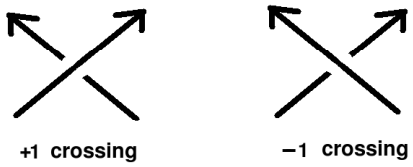

Figure 2.3: Crossing are assigned ($+1$ or -1) according to a right hand rule

Because a Θ-SAP in class I can be mapped to a unique Θ-SAP in class II (and vice versa) by the bijective symmetry map [61]

$$f((x, y, z)) = (-x, y, z),$$

the cardinality of these two classes are equal. Because f is a bijection between class I and class II Θ-SAPs, one needs only to sample from one of the two classes [60]. It is also proved in [60] that this symmetry map preserves knot type in the case of achiral knots, and reverses chirality in the case of chiral knots. An example of the application of this symmetry map is shown in Figure 2.4.

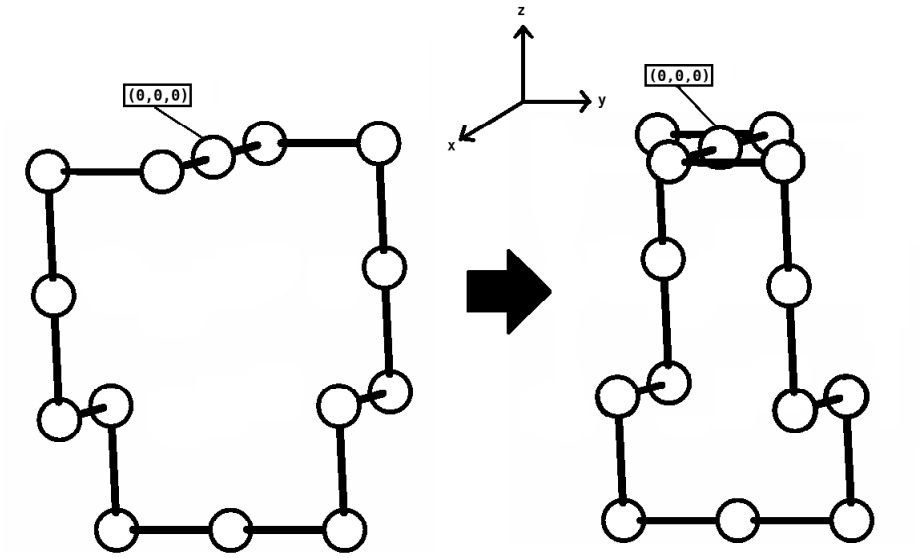

Figure 2.4: Example of the symmetry map between class I and class II Θ-SAPs

2.1.2 Strand passage in Θ-SAPs

A strand passage can be attempted on a Θ-SAP as follows:

Definition 2.1.3. *A strand passage in a Θ-SAP is defined as the process of removing the Θ-structure and replacing it with an after-strand passage structure (referred to here as η) consisting of vertices $V(\eta)$ and edges $E(\eta)$ defined by:*

$$\begin{aligned} V(\eta) &= \{(-1,0,0),(-1,0,-1),(-1,0,-2),(0,0,-2),(1,0,-2),(1,0,-1),\\ &\quad (1,0,0),(0,-1,-2),(0,-1,-1),(0,0,-1),(0,1,-1),(0,1,-2)\}, \end{aligned}$$

and

$$\begin{aligned} E(\eta) &= \{\{(1,0,0),(1,0,-1)\},\{(1,0,-1),(1,0,-2)\},\{(1,0,-2),(0,0,-2)\}\\ &\quad \{(0,0,-2),(-1,0,-2)\},\{(-1,0,-2),(-1,0,-1)\},\{(-1,0,-1),(-1,0,0)\},\\ &\quad \{(0,-1,-2),(0,-1,-1)\},\{(0,-1,-1),(0,0,-1)\},\{(0,0,-1),(0,1,-1)\},\\ &\quad \{(0,1,-1),(0,1,-2)\}\}. \end{aligned}$$

A strand passage in a Θ-SAP ω is said to be "successful" if the polygon that results from replacing the Θ-structure in ω with the η-structure is still a self-avoiding polygon. A "failed" strand passage is said to occur if the resulting after-strand passage polygon is not self-avoiding. Figure 2.5 provides an illustration of a Θ-SAP before and after a successful strand passage.

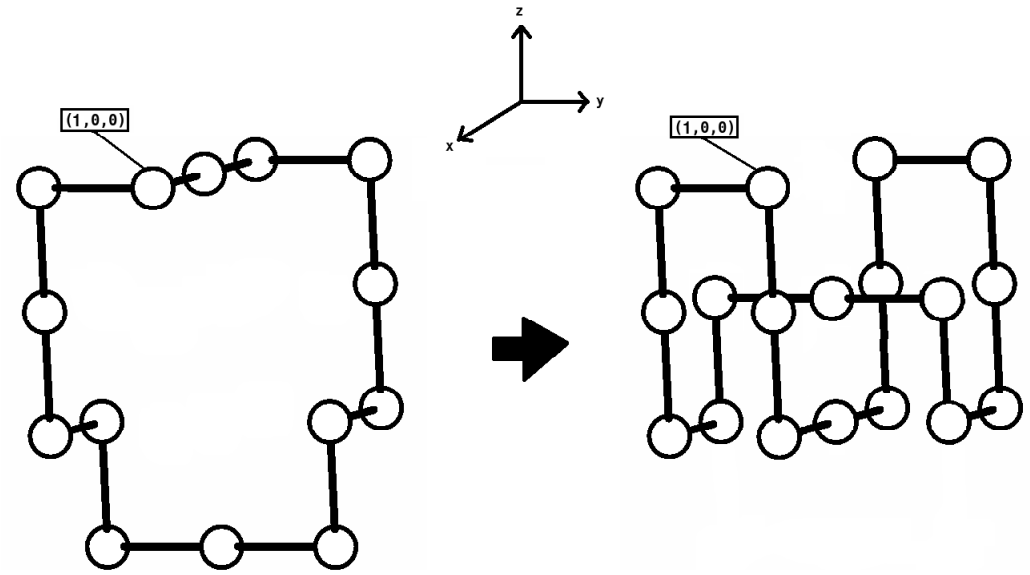

Figure 2.5: Example of a successful strand passage in a Θ-SAP

2.1.3 Theoretical Results for Θ-SAPs

This section reviews some theoretical results for Θ-SAPs; these results will be useful later for obtaining results for the ILSP model.

For the remainder of this work, all Θ-SAPs discussed will be class II Θ-SAPs (Θ^--SAPs); this leads to the following notation:

Definition 2.1.4. *For each knot type K, define $\mathcal{P}^{\Theta}_{2n}(K)$ to be the set of all class II Θ-SAPs with length $2n$ and knot type K. Let $p^{\Theta}_{2n}(K)$ be the number of class II Θ-SAPs with length $2n$ and knot type K up to translation. Define $\mathcal{P}^{\Theta}(K) := \bigcup_{n=1}^{\infty} \mathcal{P}^{\Theta}_{2n}(K)$ to be the set of all class II Θ-SAPs with knot type K.*

Definition 2.1.5. *Given any knot type K, define $n(K)$ to be the smallest number of edges for which a SAP can have knot type K, and define $n^{\Theta}(K)$ to be the smallest number of edges for which a Θ-SAP can have knot type K.*

How does $p^{\Theta}_{2n}(\phi)$ grow as $n \to \infty$? This question was answered by Szafron in [60]:

Lemma 2.1.6 ([60])**.** *For any fixed knot-type K and for all natural numbers $n \geq n^{\Theta}(K)$,*

$$p^{\Theta}_{n}(K) \leq n p_n(K). \tag{2.1}$$

Lemma 2.1.7 ([60]). *For any fixed knot-type K and for all natural numbers $n \geq \max\{n^{\Theta}(K), n(K)+14\}$,*

$$\frac{1}{2}p_{n-14}(K) \leq p_n^{\Theta}(K). \tag{2.2}$$

Theorem 2.1.8 ([60]). *The exponential growth rate for $p_{2n}^{\Theta}(\phi)$ is:*

$$\lim_{n\to\infty} \frac{1}{2n} \log p_{2n}^{\Theta}(\phi) = \kappa_{\phi}. \tag{2.3}$$

Proof. Combining Lemmas 2.1.6 and 2.1.7, for all $n \geq 9$, we get the inequality:

$$\frac{1}{2}p_{2n-14}(\phi) \leq p_{2n}^{\Theta}(\phi) \leq 2np_{2n}(\phi).$$

Taking logarithms, dividing by $2n$ and taking the limit as $n \to \infty$ yields the following:

$$\lim_{n\to\infty} \frac{1}{2n} \log p_{2n-14}(\phi) \leq \lim_{n\to\infty} \frac{1}{2n} \log p_{2n}^{\Theta}(\phi) \leq \lim_{n\to\infty} \frac{1}{2n} \log(2np_{2n}(\phi)),$$

which can be written as:

$$\kappa_{\phi} \leq \frac{1}{2n} \log p_{2n}^{\Theta}(\phi) \leq \kappa_{\phi}.$$

□

The implication of Theorem 2.1.8 is that the number of $2n$-edge unknotted Θ-SAPs grows at the same exponential rate as the number of $2n$-edge unknotted SAPs.

Because it is not known whether $k_K = \kappa_K$ (refer back to Equation 1.8) for any knot type K other than the unknot, it is only possible to bound the growth rate of $p_{2n}^{\Theta}(K)$ as follows:

Definition 2.1.9. *Define the bounds on the exponential growth rate of $p_{2n}^{\Theta}(K)$ to be:*

$$k_K^{\Theta} := \liminf_{n\to\infty} \frac{1}{2n} \log p_{2n}^{\Theta}(K) \leq \limsup_{n\to\infty} \frac{1}{2n} \log p_{2n}^{\Theta}(K) =: \kappa_K^{\Theta}. \tag{2.4}$$

Theorem 2.1.10 ([60]). *For each non-trivial knot type K,*

$$\kappa_{\phi} \leq k_K = k_K^{\Theta} \leq \kappa_K^{\Theta} = \kappa_K < \kappa. \tag{2.5}$$

It is important to define and bound all of these growth rates, as they play important roles in algorithms for generating random SAPs and Θ-SAPs.

2.2 Quantities of Interest

Now that the appropriate theory for the LSP Model has been provided, it is time to review definitions for quantities relating to knotting and compactness that are of particular interest for addressing Problems 1 and 2 in the good solvent case.

2.2.1 Probability of Knot Types and Knotting

Given a random SAP of length $2n$, a natural question one might ask is "What is the probability that this SAP is knotted?" Given a randomly (uniformly) selected SAP ω of length $2n$, denote the probability of this SAP having knot type K by

$$\rho_{2n}(K) := \Pr(k(\omega) = K \,|\, |\omega| = 2n) = \frac{p_{2n}(K)}{p_{2n}}, \tag{2.6}$$

and denote the probability of this SAP being knotted by

$$\rho_{2n}(\bar{\phi}) := \Pr(k(\omega) \neq \phi \,|\, |\omega| = 2n) = 1 - \frac{p_{2n}(\phi)}{p_{2n}} = 1 - \rho_{2n}(\phi). \tag{2.7}$$

Because the exponential growth rate of $p_{2n}(\phi)$ (*i.e.* κ_ϕ) is smaller than that of p_{2n} (*i.e.* κ), the probability of a randomly chosen SAP with length $2n$ being knotted tends to 1 as $n \to \infty$. This statement can be generalized; namely, because the exponential growth rate of $p_{2n}(K)$ (bounded above by κ_K) is smaller than κ, the probability of a randomly chosen SAP with length $2n$ having knot type K tends to 0 as $n \to \infty$ [57].

2.2.2 Probabilities Relating to Strand Passage in a Θ-SAP

A primary question relating to strand passage in Θ-SAPs is "Given a successful strand passage in a length $2n$ Θ-SAP with knot type K, what is the probability that one will end up with a SAP with knot type K'?" Before this question can be addressed, some further notation needs to be introduced. Recall from Section 2.1.3 that the term Θ-SAP now refers only to class II Θ-SAPs (*i.e.* Θ^--SAPs).

Definition 2.2.1. *Define the set of length $2n$ Θ-SAPs with knot type K for which strand passage is successful to be $\mathcal{P}^{\Theta}_{2n}(s|K)$. Denote the number of such successful strand passage Θ-SAPs to be $p^{\Theta}_{2n}(s|K)$.*

Given a knot type K and any natural number $n \geq n^{\Theta}_K/2$, the probability of a successful strand passage in a randomly chosen Θ-SAP with length $2n$ and knot type K is denoted

$$\rho^{\Theta}_{2n}(s|K) := \frac{p^{\Theta}_{2n}(s|K)}{p^{\Theta}_{2n}(K)}. \tag{2.8}$$

Definition 2.2.2. *Define $\mathcal{K}^{\Theta}(K)$ to be the set of all knot types that can result from a single strand passage in a Θ-SAP with knot type K.*

Definition 2.2.3. *Given a knot type K, a natural number $n \geq n_K^\Theta/2$, and another knot type $K' \in \mathcal{K}^\Theta(K)$, define $\mathcal{P}_{2n}^\Theta(K'|K,s)$ to be the set of all Θ-SAPs ω in $\mathcal{P}_{2n}^\Theta(s|K)$ that have knot type K' after the Θ-structure in ω is replaced with the η-structure. Denote the number of such Θ-SAPs by $p_{2n}^\Theta(K'|K,s)$.*

Given a successful strand passage in a length $2n$ ($n \geq n_K^\Theta/2$) Θ-SAP with knot type K, the probability that the resulting SAP will have knot type $K' \in \mathcal{K}^\Theta(K)$ is denoted

$$\rho_{2n}^\Theta(K \to K') := \frac{p_{2n}^\Theta(K'|K,s)}{p_{2n}^\Theta(s|K)}. \tag{2.9}$$

Another important question is "How does the strand passage probabilities defined in Equations 2.8 and 2.9 behave as $n \to \infty$?" The following limit, if it exists, is referred to as the *limiting successful strand passage probability*:

$$\rho^\Theta(s|K) := \lim_{n\to\infty} \rho_{2n}^\Theta(s|K), \tag{2.10}$$

and for each $K' \in \mathcal{K}^\Theta(K)$, the following limits, if they exist, are referred to as *limiting knot-transition probabilities*:

$$\rho^\Theta(K \to K') := \lim_{n\to\infty} \rho_{2n}^\Theta(K \to K'). \tag{2.11}$$

The existence of these limiting probabilities is an open question.

How do $p_{2n}^\Theta(s|K)$ and $p_{2n}^\Theta(K'|K,s)$ grow with n? For the case of the unknot, Szafron [61] proved the following:

Theorem 2.2.4 ([61])**.**

$$\kappa_{s|\phi}^\Theta := \lim_{n\to\infty} \frac{1}{2n} \log p_{2n}^\Theta(s|\phi) = \kappa_\phi. \tag{2.12}$$

Theorem 2.2.5 ([61])**.** *For any knot type $K \in \mathcal{K}^\Theta(\phi)$,*

$$\kappa_{K|\phi,s}^\Theta := \lim_{n\to\infty} \frac{1}{2n} \log p_{2n}^\Theta(K|\phi,s) = \kappa_\phi. \tag{2.13}$$

Because the exponential growth rates of $p_{2n}^\Theta(\phi)$, $p_{2n}^\Theta(s|\phi)$, and $p_{2n}^\Theta(K|\phi,s)$ are identical to that of $p_{2n}(\phi)$, it is conjectured [61] that the scaling forms for $p_{2n}^\Theta(\phi)$, $p_{2n}^\Theta(s|\phi)$ and $p_{2n}^\Theta(K|\phi,s)$ are:

$$A_\phi^\Theta (2n)^{\alpha_\phi^\Theta - 3} \mu_\phi^{2n} \left(1 + \frac{B_\phi^\Theta}{(2n)^{\Delta_\phi^\Theta}} + O(n^{-1})\right), \tag{2.14}$$

$$A^{\Theta}_{s|\phi}(2n)^{\alpha^{\Theta}_{s|\phi}-3}\mu_{\phi}^{2n}\left(1+\frac{B^{\Theta}_{s|\phi}}{(2n)^{\Delta^{\Theta}_{s|\phi}}}+O(n^{-1})\right), \tag{2.15}$$

and

$$A^{\Theta}_{K|s,\phi}(2n)^{\alpha^{\Theta}_{K|s,\phi}-3}\mu_{\phi}^{2n}\left(1+\frac{B^{\Theta}_{K|s,\phi}}{(2n)^{\Delta^{\Theta}_{K|s,\phi}}}+O(n^{-1})\right). \tag{2.16}$$

Assuming the scaling forms are valid for p^{Θ}_{2n}, $p^{\Theta}_{2n}(s|\phi)$ and $p^{\Theta}_{2n}(K|\phi,s)$, then the scaling form for $\rho^{\Theta}_{2n}(s|\phi)$ is

$$\frac{A^{\Theta}_{s|\phi}}{A^{\Theta}_{\phi}}\left((2n)^{\alpha^{\Theta}_{s|\phi}-\alpha^{\Theta}_{\phi}}\right)\left(\frac{1+\frac{B^{\Theta}_{s|\phi}}{(2n)^{\Delta^{\Theta}_{s|\phi}}}+O(n^{-1})}{1+\frac{B^{\Theta}_{\phi}}{(2n)^{\Delta^{\Theta}_{\phi}}}+O(n^{-1})}\right), \tag{2.17}$$

and the scaling form for $\rho^{\Theta}_{2n}(\phi\rightarrow K|s)$ is

$$\frac{A^{\Theta}_{K|s,\phi}}{A^{\Theta}_{s|\phi}}\left((2n)^{\alpha^{\Theta}_{K|s,\phi}-\alpha^{\Theta}_{s|\phi}}\right)\left(\frac{1+\frac{B^{\Theta}_{K|s,\phi}}{(2n)^{\Delta^{\Theta}_{K|s,\phi}}}+O(n^{-1})}{1+\frac{B^{\Theta}_{s|\phi}}{(2n)^{\Delta^{\Theta}_{s|\phi}}}+O(n^{-1})}\right). \tag{2.18}$$

Because it is expected that $\alpha^{\Theta}_{K|s,\phi}=\alpha^{\Theta}_{s|\phi}=\alpha^{\Theta}_{\phi}$ [61], then it is expected that as $n\rightarrow\infty$

$$\rho^{\Theta}_{2n}(s|\phi)=\frac{A^{\Theta}_{s|\phi}}{A^{\Theta}_{\phi}}+\frac{B'^{\Theta}_{s|\phi}}{(2n)^{\Delta'^{\Theta}_{s|\phi}}}+O(n_{s|\phi}), \tag{2.19}$$

and

$$\rho^{\Theta}_{2n}(\phi\rightarrow K)=\frac{A^{\Theta}_{K|s,\phi}}{A^{\Theta}_{s|\phi}}+\frac{B'^{\Theta}_{K|s,\phi}}{(2n)^{\Delta'^{\Theta}_{K|s,\phi}}}+O(n_{K|s,\phi}), \tag{2.20}$$

where

$$n_{s|\phi}=\min\left\{n^{-1},\max\{n^{-\Delta^{\Theta}_{\phi}},n^{-\Delta^{\Theta}_{s|\phi}}\}\right\} \tag{2.21}$$

and

$$n_{K|s,\phi}=\min\left\{n^{-1},\max\{n^{-\Delta^{\Theta}_{s|\phi}},n^{-\Delta^{\Theta}_{K|s,\phi}}\}\right\}. \tag{2.22}$$

2.2.3 Mean Square Radius of Gyration

An interesting quantity relating to ring polymers is how much space a particular ring polymer occupies. One would expect that the more compact a random ring polymer is, the higher the chance is that polymer will be knotted. One such measurement of compactness is the mean square radius of gyration, defined in our model as follows:

Definition 2.2.6. *Given a SAP ω with vertices $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$, where $\mathbf{v}_i = (x_i, y_i, z_i)$, the* square radius of gyration of ω *is defined to be:*

$$r^2(\omega) := \frac{1}{n} \sum_{i=1}^{n} \left([x_i - \bar{x}]^2 + [y_i - \bar{y}]^2 + [z_i - \bar{z}]^2 \right), \tag{2.23}$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$, and $\bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i$.

Given a set of SAPs $\mathcal{S}$, where $0 < |\mathcal{S}| < \infty$, define the *mean square radius of gyration of the elements in* $\mathcal{S}$ to be:

$$\bar{r}^2(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{\omega \in \mathcal{S}} r^2(\omega). \tag{2.24}$$

If this set $\mathcal{S}$ is $\mathcal{P}_{2n}$, then $\bar{r}^2(\mathcal{P}_{2n})$ is expected to scale asymptotically like [35]:

$$\bar{r}^2(\mathcal{P}_{2n}) \sim A(2n)^{2\nu}(1 + bn^{-d} + \ldots). \tag{2.25}$$

In [35], ν has been estimated to be 0.5877 ± 0.0006.

2.2.4 Contacts

Another such measure of compactness in our model is related to the number of contacts in a SAP. Given a SAP ω with vertices $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$, where $\mathbf{v}_i = (x_i, y_i, z_i)$, a *contact* is said to occur between vertices $\mathbf{v}_i$ and $\mathbf{v}_j$ if

$$|x_i - x_j| + |y_i - y_j| + |z_i - z_j| = 1,$$

and $\mathbf{v}_i$ and $\mathbf{v}_j$ are not connected together by an edge in the SAP. Examples of contacts are shown in Figure 2.6.

Define $C(\omega)$ to be the total number of contacts in ω. One would expect that the more compact ω is (*i.e.* the smaller $r^2(\omega)$ is), the larger $C(\omega)$ will be on average.

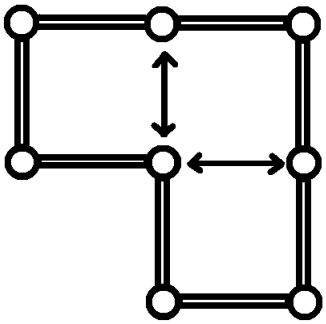

Figure 2.6: An example of a SAP with contacts. Contacts are indicated by arrows.

The reason that these compactness and knotting quantities are interesting is because in ring polymers, these quantities are sensitive to the quantity of salt in the solution. Experimentally, it has been seen [50, 53] that as the salt concentration of the solution increases, ring polymers become more compact and more likely to be knotted. If the energy model being used in this work (presented formally in the next section) is a good way to model ring polymers in salt solution, it should produce observations of these knotting and compactness quantities that are qualitatively consistent with what is obtained in these experimental results.

2.2.5 Energy of a SAP

One of the goals of this work is to model the interactions that occur between monomers of a ring polymer based on the salt concentration of the solution. The interaction model used here was first used for SAPs by Tesi *et al.* in [64]. This model includes a short range attractive force between monomers, as well as a screened Coulomb potential between monomers where the screening can be varied to account for the effect of added salt [64]. These interactions can be approximated by a Yukawa-type potential which represents the effective ion-ion potential in a Debye-Huckel model for ions in a continuum dielectric solvent [21]. The use of this model in [64] has provided results that are qualitatively consistent with experimental results obtained by Shaw and Wang in [53] and by Rybenkov *et al.* in [50].

Definition 2.2.7. *Given constants $A > 0$ (unrelated to the A defined in Equation 1.10), $\zeta > 0$, $T > 0$, $v < 0$ and a SAP ω consisting of vertices $\mathbf{v}_1, \ldots, \mathbf{v}_n$ and edges defined by the set $E(\omega)$, define the* potential energy *of ω to be*

$$U_{\zeta,A,T,v}(\omega) = \sum_{i<j\leq n} I((\mathbf{v}_i, \mathbf{v}_j) \notin E(\omega)) \left[I(r_{ij} = 1) k_B T v + \frac{A e^{-\zeta r_{ij}(\omega)}}{r_{ij}(\omega)} \right], \tag{2.26}$$

where $r_{ij}(\omega)$ is the euclidean distance between $\mathbf{v}_i$ and $\mathbf{v}_j$ in lattice units, I is the indicator function and k_B is the Boltzmann constant.

The potential energy of a SAP ω can be simplified into two parts as follows:

$$U_{\zeta,A,T,v}(\omega) = C(\omega)k_BTv + D_{A,\zeta}(\omega), \tag{2.27}$$

where

$$C(\omega) = \sum_{i<j\leq n} I((\mathbf{v}_i, \mathbf{v}_j) \notin E(\omega))I(r_{ij} = 1) \geq 0 \tag{2.28}$$

and

$$D_{A,\zeta}(\omega) := \sum_{i<j\leq n} I((\mathbf{v}_i, \mathbf{v}_j) \notin E(\omega))\frac{Ae^{-\zeta r_{ij}(\omega)}}{r_{ij}(\omega)} \geq 0. \tag{2.29}$$

This simplification is useful in proofs that will occur in later chapters. It should also be noted that in a simulation it is never actually necessary to compute $U(\omega)$, one only needs to calculate

$$U(\omega)/k_BT = C(\omega)v + \sum_{i<j\leq n}\left[I((\mathbf{v}_i, \mathbf{v}_j) \notin E(\omega))\left(\frac{A}{k_BT} \times \frac{e^{-\zeta r_{ij}(\omega)}}{r_{ij}(\omega)}\right)\right]. \tag{2.30}$$

The parameter ζ^{-1} in this model is called the *Debye length* [21], measured in lattice units; its value reflects the ionic strength of the solution [64] where larger values of ζ correspond to a higher salt concentration in the model. For a *1-1 electrolyte* (such as NaCl), the Debye length can be related to the concentration of the solution as follows [33]:

$$\zeta^{-1} = \left(\frac{\epsilon k_BT}{e^2N_a2c}\right)^{\frac{1}{2}}, \tag{2.31}$$

where ϵ is the *absolute permittivity* of the solution, e is the charge of an electron, N_a is *Avogadro's number*, and c is the concentration of salt in moles per cubic meter.

The parameter A in this model is connected to the charge density along the polymer chain [64] (this parameter is not related to the amplitude of the scaling form specified in Equation 1.10). In [64], the authors choose 3 different values for A, namely those defined by $A/k_BT = 0.01$, 0.1, and 1; their results suggest that the knotting probabilities are not highly sensitive to the choice of A [64].

The parameter v in this model is used as a short range attactive force between monomers; this force only exists for non-bonded monomers that are unit distance apart (*i.e.* contacts). v is chosen to be -0.26 because for the simple cubic lattice, it is known that this value corresponds to a poor

solvent regime where the knot probability is higher and easier to study [14, 64]. The case where $A = 0$ in this model, *i.e.* where

$$U_{\zeta,A,T,v}(\omega) = C(\omega)k_BTv, \tag{2.32}$$

has been well studied in [14, 65, 66]. The results in these articles show that for values of v that are less than some number v_c (related to the *collapse transition temperature*, also known as the *θ-temperature*), self avoiding polygons in the model tend to 'collapse' into a ball. The value for v_c is estimated in [65] to be -0.2782 ± 0.007. One can notice that the value of v used in this work (-0.26) is chosen to be greater than but quite close to v_c.

When considering SAPs that model ring polymers in a good solvent, the standard assumption is that all conformations of SAPs with the same length are equally likely. If one is considering ring polymers modelled by SAPs with an energy reflective of the solvent conditions as defined by Equation 2.26, it is no longer assumed that two SAPs with the same length are equally likely. In this model, two SAPs with the same length and the same energy are now equally likely. The next section will describe what kind of probability distributions are of interest relating to these model assumptions, as well as different probability distributions depending on the desired sample space.

2.3 Probability Distributions of Interest

The goal of this thesis is to generate samples of SAPs according to distributions relating to the model being used. In this work, we are interested primarily in two different sample spaces. The first sample space consists of SAPs with a fixed length n and variable knot type (*i.e* $\mathcal{P}_n$), whereas the second sample space relates to SAPs with a fixed structure (Θ) and fixed knot type K, but variable length (*i.e.* $\mathcal{P}^{\Theta}(K)$). Distributions relating to these sample spaces will be presented for both the good solvent and varying solvent models.

2.3.1 Good Solvent Model

Suppose that we are interested in sampling SAPs of a fixed length n, but are not interested in restricting its knot type. This sample space corresponds to the set of all SAPs of length n, namely, $\mathcal{P}_n$. In the good solvent model, all SAPs with the same length are considered to be equally likely [61]; thus, every SAP in $\mathcal{P}_n$ should have an equal chance of being selected. Therefore, the desired probability distribution for this case is simply a uniform distribution, where every SAP $\omega \in \mathcal{P}_n$ has a probability of $\frac{1}{p_n}$.

Now suppose that we are interested in sampling SAPs of variable length, but with a fixed structure (*i.e.* Θ-SAPs) and fixed knot type K; the sample space corresponding to this is $\mathcal{P}^{\Theta}(K)$. Again, in the good solvent model all Θ-SAPs of the same length should have an equal probability. However, the sample space is now countably infinite; thus, there is a need to come up with a distribution consisting of finite non-zero probabilities. For any $\omega \in \mathcal{P}^{\Theta}(K)$ such that $|\omega| = n$, one such choice is such that

$$\Pr(\omega) = \frac{w(n)z^n}{\sum\limits_{\omega' \in \mathcal{P}^{\Theta}(K)} w(|\omega'|)z^{|\omega'|}} =: \frac{w(n)z^n}{Q_K^{\Theta}(z,w)}, \tag{2.33}$$

where $w(n)$ is a fixed polynomial weight function of polygon length n and $0 < z < e^{-\kappa_K}$ (κ_K is defined in Equation 1.8). When $w(n) = 1$, this distribution corresponds to the standard *grand canonical ensemble* [40] from statistical mechanics. Provided that $w(n)$ is a polynomial weight function, the radius of convergence of $Q_K^{\Theta}(z,w)$ in Equation 2.33 is $e^{-\kappa_K}$.

Note that it is also possible to sample from the set of all unrooted SAPs with knot type K, such that for any $\omega \in \mathcal{P}(K)$,

$$\Pr(\omega) = \frac{w(n)z^n}{\sum\limits_{\omega' \in \mathcal{P}(K)} w(|\omega'|)z^{|\omega'|}} =: \frac{w(n)z^n}{Q_K(z,w)}; \tag{2.34}$$

this will be discussed briefly in Section 4.2.

2.3.2 Varying Solvent Model

When the model incorporates a salt solution, it is no longer assumed that two SAPs of the same length are equally likely. It is, however, assumed that two SAPs with the same length and the same energy are equally likely. If we are interested in sampling SAPs of a fixed length n where the knot type is allowed to vary, given a set of "appropriate" energy parameters $\mathcal{E} = \{\zeta, A, T, v\}$ (where by "appropriate" it is meant that $\zeta > 0$, $A \geq 0$, $T > 0$ and $\nu \leq 0$), we can sample from a Boltzmann distribution where the probability of obtaining some $\omega \in \mathcal{P}_n(K)$ is:

$$\Pr(\omega) = \frac{e^{\frac{-U_{\mathcal{E}}(\omega)}{k_B T}}}{\sum\limits_{\omega' \in \mathcal{P}_n(K)} e^{\frac{-U_{\mathcal{E}}(\omega')}{k_B T}}} =: \frac{e^{\frac{-U_{\mathcal{E}}(\omega)}{k_B T}}}{Z_n(\mathcal{E})}. \tag{2.35}$$

This probability distribution is valid because there is a finite number of states, each with a finite energy.

If we are assuming a salt solution and are interested in sampling from $\mathcal{P}^{\Theta}(K)$, we can extend the distribution defined in Equation 2.33 to also be weighted by the energy of a SAP. Given a set of appropriate energy parameters $\mathcal{E} = \{\zeta, A, T, v\}$, the resulting distribution for a length n Θ-SAP $\omega \in \mathcal{P}^{\Theta}(K)$ is:

$$\Pr(\omega) = \frac{e^{\frac{-U_{\mathcal{E}}(\omega)}{k_B T}} w(n) z^n}{\sum_{\omega' \in \mathcal{P}^{\Theta}(K)} e^{\frac{-U_{\mathcal{E}}(\omega')}{k_B T}} w(|\omega'|) z^{|\omega'|}} =: \frac{e^{\frac{-U_{\mathcal{E}}(\omega)}{k_B T}} w(n) z^n}{Q^{\Theta}_{K,\mathcal{E}}(z, w)}, \tag{2.36}$$

where $w(n)$ is a polynomial weight function of polygon length n and $0 < z < z_c(\mathcal{E})$. $z_c(\mathcal{E})$ will vary depending on the energy parameters being selected; in Section 5.3.1 it is proved that $z_c(\mathcal{E})$ will be positive for any set of appropriate energy parameters. Equation 2.36 defines the distribution for the ILSP model.

2.4 Chapter Summary

This chapter reviewed the Local Strand Passage (LSP) Model for SAPs in the simple cubic lattice, developed by Szafron in [60] and [61]. The SAPs in this model (referred to as Θ-SAPs) contain a fixed structure Θ that represents two strands of the SAP being brought close together for the purpose of a strand passage. A strand passage can be attempted on a Θ-SAP by replacing the Θ-structure with an alternate structure (referred to here as the η structure); this replacement procedure models the strand passage action of type II topoisomerase enzymes. Some key theoretical results pertaining to the LSP model were also reviewed; the LSP model and these theoretical results form the basis for the new ILSP model in this thesis.

Some observable quantities of interest for SAPs and Θ-SAPs were defined in this chapter. These quantities provide a measure of compactness and knottedness of a SAP, and it is of interest to see how these quantities change with different solvent conditions.

Equation 2.26 describes a potential energy that approximates the interactions that occur between monomers of a ring polymer based on the salt concentration of the solution. This potential energy includes a short range attractive force between monomers, as well as a screened Coulomb potential between monomers that can be varied to account for the effect of salt in the model.

In the work presented here, we are interested in sampling from two primary sample spaces. The first sample space consists of all SAPs of a fixed length n, whereas the second sample space relates to Θ-SAPs with a fixed knot type K. Probability distributions relating to these sample spaces were

presented in Section 2.3 for the good solvent and varying solvent models.

Now that the models for the assumptions of good and varying solvents have been defined, the next step is to discuss the theory and methods relating to Markov Chains. This methodology is essential, as it holds the key to generating samples of random SAPs and Θ-SAPs for the different models and probability distributions described here.

Chapter 3

Markov Chain Theory

The following chapter reviews some basic Markov Chain theory, as well as different Monte Carlo methods and how to analyze their output. Markov chains can be extremely useful in generating samples from distributions that can be rigorously defined but are difficult to sample from directly.

The theory of Markov chains will be needed to develop the CMC Θ-BFACF algorithm for the ILSP model. The theory and methods described in this chapter for analyzing data from a Markov Chain or Composite Markov Chain are largely those reviewed and used previously by Szafron in [61]; however, they are being reviewed again because I wrote my own versions of these analysis programs, and also for referencing in future sections.

3.1 Basic Notation and Theory

Unless otherwise stated, the following terminology and theory is based on [31].

A *stochastic process* is a family of random variables X_t defined on a probability space $(\Omega,\mathcal{F},P)$, where t is defined on an index set Λ which can be discrete or continuous. Such a process is denoted by $\{X_t, t \in \Lambda\}$. From now on, assume that this index set is the set of non-negative integers, defined to be T. The values that X_t can assume are called *states*; the set of all possible states is called the *state space* (denoted by $\mathcal{S}$).

Definition 3.1.1. *A stochastic process has the* Markov property *if the following is true: given that at time t the random variable X_t is state x_t (here it is assumed that X_t takes on real values), then for all $s > t$,*

$$Pr(a < X_s \leq b | X_1 = x_1, \ldots, X_t = x_t) = Pr(a < X_s \leq b | X_t = x_t). \tag{3.1}$$

Define the *one-step transition probability* to go to state y at time $t+1$ given that the state at time t is x to be

$$P_{xy}^t = \Pr(X_{t+1} = y | X_t = x). \tag{3.2}$$

If P_{xy}^t is indepedent of t, then the Markov chain is *time homogeneous* and P_{xy}^t is denoted by P_{xy}. These time homogeneous probabilities P_{xy} for all possible states can be represented by the *one-step transition probability matrix* $\mathbb{P} = (P_{xy})_{x,y\in\mathcal{S}}$.

Define the *n-step transition probability* to go to state y at time $t+n$ given that the state at time t is x to be

$$P_{xy}^{(n)} = \Pr(X_{t+n} = y | X_t = x). \tag{3.3}$$

Define $(P_{xy}^{(n)})_{x,y\in\mathcal{S}} := \mathbb{P}^{(n)}$ to be the *n-step transition probability matrix* (note that $\mathbb{P}^{(n)} = \mathbb{P}^n$ using the usual matrix power). A time-homogeneous Markov chain can be completely specified by $\mathbb{P}$.

Definition 3.1.2. *A Markov chain* $\{X_t, t \in T\}$ *is said to be* stationary *if, for arbitrary* $t_1, t_2, \ldots, t_n \in T$ *and any* $n \in \mathbb{Z}^+$, *the joint distributions of* $(X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_n+h})$ *and* $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$ *are the same for all* $h > 0$.

Definition 3.1.3. *State* j *is said to be* accessible *from state* i *if for some integer* $n \geq 0$, $P_{ij}^{(n)} > 0$*: i.e., state* j *is accessible from state* i *if there is positive probability that in a finite number of transitions state* j *can be reached starting from state* i. *Two states* i *and* j, *each accessible to each other, are said to* communicate.

Definition 3.1.4. *A Markov chain is defined to be* irreducible *if for all* $x, y \in \mathcal{S}$, x *and* y *communicate.*

Definition 3.1.5. *Define the* period *of a state* i, *written as* $d(i)$, *to be the greatest common divisor of all integers* $n \geq 1$ *for which* $P_{ii}^{(n)} > 0$. *A Markov chain is called* aperiodic *if every state* $i \in \mathcal{S}$ *has* $d(i) = 1$.

Definition 3.1.6. *Let* $f_{xy}^{(i)}$ *denote the probability that it will take* i *transitions to first reach state* y *from state* x. *A state* x *is defined to be* recurrent *if and only if*

$$\sum_{i=1}^{\infty} f_{xx}^{(i)} = 1. \tag{3.4}$$

Another way of determining recurrence is as follows. A state x is recurrent if and only if

$$\sum_{i=1}^{\infty} P_{xx}^{(i)} = \infty. \tag{3.5}$$

A Markov chain is said to be recurrent if all states $x \in \mathcal{S}$ are recurrent. If states i and j communicate, and i is recurrent, then j is recurrent as well. Therefore, an irreducible Markov chain

with one recurrent state implies that all other states are recurrent, and thus the Markov Chain is recurrent.

A Markov chain is said to be *positive recurrent* if it is recurrent and there exists some $x \in \mathcal{S}$ where $\lim_{n\to\infty} P_{xx}^{(n)} > 0$.

Definition 3.1.7. *A Markov chain is said to be* reversible *if it is positive recurrent, irreducible, and* $\forall\, x, y \in \mathcal{S}, \exists$ *probabilities* π_x *and* π_y *such that* $\pi_x P_{xy} = P_{yx}\pi_y$.

Definition 3.1.8. *In this work a Markov chain is referred to as being* ergodic *if it is positive recurrent, irreducible and aperiodic.*

Definition 3.1.9. *A set of probabilities* $\pi := \{\pi_x\}_{x\in\mathcal{S}}$ *is called a* stationary distribution *(or* equilibrium distribution*) of a Markov Chain* $\{X_t, t \in T\}$ *if* $\forall\, x \in \mathcal{S}$,

$$\pi_x \geq 0, \sum_{x\in\mathcal{S}} \pi_x = 1, \text{ and } \sum_{y\in\mathcal{S}} \pi_y P_{yx} = \pi_x. \tag{3.6}$$

Theorem 3.1.10 ([7])**.** *If a Markov chain* $\{X_t, t \in T\}$ *defined by the transition probability matrix* $\mathbb{P}$ *is irreducible and positive recurrent, then a unique equilibrium distribution* π *exists and* $\pi_x > 0$ *for all* x*. If* $\mathbb{P}$ *is aperiodic, then* $\lim_{n\to\infty} P_{xy}^{(n)} = \pi_y$.

Thus, if a chain is reversible, then there exists a unique equilibrium distribution for the chain.

The following theorem from [22] provides a nice criterion for verifying the equilibrium distribution of a Markov chain:

Theorem 3.1.11 ([22])**.** *For an irreducible Markov chain, if there exists a distribution* $\pi := \{\pi_x\}_{x\in\mathcal{S}}$ *such that* $0 \leq \pi_x \leq 1$, $\sum_{x\in\mathcal{S}} \pi_x = 1$ *and* $\pi_x P_{xy} = P_{yx}\pi_y \;\; \forall\, x, y \in \mathcal{S}$*, then the chain is reversible with unique equilibrium distribution* π.

3.2 Markov Chain Monte Carlo Simulations

Markov Chain Monte Carlo (MCMC) methods are Monte Carlo simulations using Markov chains. Suppose one is interested in sampling from a discrete set $\mathcal{S}$ according to some probability mass function $\pi := \{\pi_x\}_{x\in\mathcal{S}}$. Suppose one can define a Markov chain on $\mathcal{S}$ whose unique equilibrium distribution is π. Even though this Markov chain produces a series of correlated states, it is possible that the correlation between two states in the chain can be considered negligible if they are separated by a 'large enough' amount of time steps. If this is true, then given a large enough sample one can make inferences about the desired distribution π based on the sample from the Markov chain.

When using MCMC methods, there are many questions that need to be addressed in order to make reasonable inferences. For example, how many time steps must pass before a Markov chain reaches its desired equilibrium distribution? Because the states in a Markov chain are correlated, how many time steps must pass before the correlation between two states in the chain can be considered negligible? These are difficult questions that cannot necessarily be answered exactly; later sections will present methods designed to address these and other questions related to MCMC methods.

3.2.1 Composite Markov Chains

Composite Markov Chains (*CMCs*), originally called *Metropolis-coupled Markov Chain Monte Carlo*, was introduced by Geyer in 1991 [20]. The concept of CMCs have also been referred to as *Multiple Markov Chain* Sampling [66], *Parallel Tempering* [25], and *Exchange Monte Carlo* [28]. A CMC involves several Markov chains run in parallel, with periodic swapping of states attempted between different chains in the system. This swapping allows for the possibility to immediately introduce a completely different state into a particular chain in the CMC. This is particularly helpful in chains where there is a potential for a state to get stuck in a local state for a long period of time. Such 'local equilibriums' can provide the appearance that the Markov chain has converged globally; this opens up the possibility to make misleading inferences. Because this swapping introduces dependence between the Markov chains, each chain by itself is no longer 'Markov'; however, the whole system is a Markov chain - hence the term 'composite'.

This technique has been used by Orlandini in [45], as well as Szafron in [60] and [61] to study self-avoiding polygons. The following terminology and concepts are adapted from [45].

Suppose we have M Markov chains (on the same state space $\mathcal{S}$) being run in parallel, where $\pi(i)$ is the equilibrium distribution of the i'th Markov chain. These equilibrium distributions should be chosen such that there is 'considerable overlap' between the distributions of adjacent chains. Suppose these chains have been run in parallel for a specified number of time steps (say, t^*). Choose two chains (i and j) with probability p^*_{ij} according to some probability distribution over all pairs of chains where $p^*_{ij} = p^*_{ji}$. Suppose x is the state in chain i and y is the state in chain j; these states will be *swapped* between the two chains with probability

$$r(i,j) = \min\left(1, \frac{\pi_y(i)\pi_x(j)}{\pi_y(j)\pi_x(i)}\right), \tag{3.7}$$

where $r(i,j)$ is referred to as the swap probability between chains i and j. One time step in the

CMC consists of either one non-swap move on each of the M chains in the CMC (a move or time step in parallel), or an attempted swap between two chains.

A formal definition of the CMC is as follows:

Definition 3.2.1 ([61]). *Given $M > 0$, state space $\mathcal{S}$, and for each $i = 1, \ldots, M$, let $\{X_t(i), t \in T\}$ be an ergodic Markov chain on $\mathcal{S}$ with one step transition probabilities defined by $\{P_{xy}(i)\}_{x,y\in\mathcal{S}}$ with its equilibrium distribution given by $\pi(i) = \{\pi_x(i)\}_{x\in\mathcal{S}}$. Suppose t^* is some positive integer and for each i and j, p^*_{ij} is chosen to satisfy $p^*_{ij} = p^*_{ji}$ and $\sum_{i,j\leq M} p^*_{ij} = 1$. Then define the composite chain $\{Y_t, t \in T\}$ with $Y_t = (X_t(1), \ldots, X_t(M)) \in \mathcal{S}^M$ to be the stochastic process on the state space $\mathcal{S}^M$ with one step transition probabilities specified by*

$$P_{xy} = \begin{cases} \prod_{i=1}^{M} P_{x_i y_i}(i), & \text{if } t \bmod (t^*+1) \neq 0, \\ p^*_{ij} r(i,j), & \text{if } t \bmod (t^*+1) = 0, y = x_s(i,j), \\ 1 - \sum_{i,j} p^*_{ij} r(i,j), & \text{if } t \bmod (t^*+1) = 0, y = x, \\ 0, & \text{otherwise,} \end{cases} \tag{3.8}$$

where for $x = (x(1), \ldots, x(i), \ldots, x(j), \ldots, x(M)) \in \mathcal{S}^M$,

$$x_s(i,j) = (x(1), \ldots, x(i-1), x(j), x(i+1), \ldots, x(j-1), x(i), x(j+1), \ldots, x(M)) \in \mathcal{S}^M, \tag{3.9}$$

and

$$r(i,j) = \min\left(1, \frac{\pi_y(i)\pi_x(j)}{\pi_y(j)\pi_x(i)}\right). \tag{3.10}$$

If each chain in the CMC is ergodic, then so is the CMC, and the unique equilibrium distribution of the CMC is the product of the equilibrium distributions of the separate chains. The swapping does not change the equilibrium distributions of the whole process, but is expected to make the convergence of each chain to its equilibrium distribution faster than if they were run independently [45].

There are several issues that need to be addressed when running a CMC:

What should the distributions be for the M Markov chains? What is a suitable choice for M? The answers to these questions vary depending on the type of distributions being considered, most notably in how fast these distributions converge. In the first chain you will want a distribution that is known to converge quite rapidly to its equilibrium distribution, while in the last chain you want a distribution that might take a long (but not too long) time to converge [45]. There is an algorithm described in [45] which describes the principles of how to choose a suitable number of chains M and distributions of chains 2 to $M-1$ based on the distributions of chains 1 and M. This procedure

orders the chains so that the number of time steps it takes for a chain to converge increases as i increases from 1 to M.

Another question one might ask is with respect to choosing the values of p^*_{ij} (i.e. the probabilities for attempting a swap between chains i and j), and how often one needs to attempt a swap (choice of t^*). Recall that the distributions for the M chains are chosen so that there is considerable overlap in the distributions of adjacent chains. If $p^*_{ij} > 0$ for all possible pairs i and j, then it is possible to swap two chains that have very different distributions, possibly causing a huge change in the states of those chains. However, such changes are not likely to be accepted often because there is minimal overlap between their distributions, possibly making it not worthwhile to have. In the work presented here, swaps are only attempted between adjacent chains because there is the most overlap between distributions and the best chance of a successful swap. As for the choice of t^*, there is no simple way to determine the optimal time between swaps; one way to get an estimate would be to run several CMCs with different swap rates and see how the time it takes for the system to converge to equilibrium differs. The swap rates used here were taken from comparable CMCs used in [60] and [61].

3.2.2 Convergence to the Equilibrium Distribution

As mentioned in the preamble of this section, an important question that needs to be addressed by Markov Chain theory is determining if and when the Markov chain has reached its equilibrium distribution. Because a Markov Chain is rarely started from its equilibrium distribution, data generated during an initial period of some length will not be reflective of the true distribution of interest. The following discussion for this topic is based on [55].

Suppose we have a stationary stochastic process $\mathcal{X} = \{X_t, t \in T\}$ with equilibrium distribution $\pi = \{\pi_x\}_{x \in \mathcal{S}}$. Define $\mathcal{H}$ to be the set of all real valued functions defined on $\mathcal{X}$; for each $f \in \mathcal{H}$, the function $f(\mathcal{X})$ is referred to as an *observable* of $\mathcal{X}$. The set $f(\mathcal{X}) := \{f(X_t), t \in T\}$ is also a stationary stochastic process [31].

Definition 3.2.2. *For a stationary stochastic process $f(\mathcal{X})$ with equilibrium distribution $\pi = \{\pi_x\}_{x \in \mathcal{S}}$, define the* mean of f with respect to π *to be:*

$$E_\pi(f) := \sum_{x \in \mathcal{S}} f(x)\pi_x. \tag{3.11}$$

Definition 3.2.3. *For a stationary stochastic process $f(\mathcal{X})$ with equilibrium distribution $\pi =$*

$\{\pi_x\}_{x\in\mathcal{S}}$, *define the* variance of f with respect to π *to be:*

$$\mathrm{var}_\pi(f) := \sum_{x\in\mathcal{S}} (f(x) - E_\pi(f))^2 \pi_x. \tag{3.12}$$

The following two definitions come from Fishman in [16]:

Definition 3.2.4 ([16])**.** *For a stationary stochastic process $f(\mathcal{X})$ with equilibrium distribution $\pi = \{\pi_x\}_{x\in\mathcal{S}}$, define the* autocovariance function *of f with respect to π, denoted $\gamma_f(h)$, to be:*

$$\gamma_f(h) := E_\pi(f(X_t)f(X_{t+h})) - (E_\pi(f))^2 = \sum_{x,y\in\mathcal{S}} f(x)\left[\pi_x P_{xy}^{(|h|)} - \pi_x\pi_y\right] f(y). \tag{3.13}$$

Definition 3.2.5 ([16])**.** *For a stationary stochastic process $f(\mathcal{X})$ with equilibrium distribution $\pi = \{\pi_x\}_{x\in\mathcal{S}}$, define the* autocorrelation function *of f with respect to π, denoted $\rho_f(h)$, to be:*

$$\rho_f(h) = \frac{\gamma_f(h)}{\gamma_f(0)}. \tag{3.14}$$

Definition 3.2.6. *Given a stationary stochastic process $f(\mathcal{X})$ started in its equilibrium distribution $\pi = \{\pi_x\}_{x\in\mathcal{S}}$, define the* exponential autocorrelation time *[55] of the observable f to be:*

$$\tau_{\exp}(f) := \limsup_{h\to\infty} \frac{h}{-\log|\rho_f(h)|}, \tag{3.15}$$

where $\rho_f(h)$ is the autocorrelation function with respect to π.

Definition 3.2.7. *If $\mathcal{H}$ is the set of all observable functions of an ergodic Markov chain, define the* exponential autocorrelation time *of the Markov chain to be:*

$$\tau_{\exp} := \sup_{f\in\mathcal{H}} \tau_{\exp}(f). \tag{3.16}$$

Supposing now that the starting state of the Markov chain is not from the equilibrium distribution π, Sokal showed in [55] that the convergence of the chain to the equilibrium distribution is bounded above by $\tau_{\exp}$. If we can estimate the value $\tau_{\exp}(f)$ for the function f that takes the longest time to reach its equilibrium distribution in a particular ergodic Markov chain, we can be confident that every observable $f \in \mathcal{H}$ has reached its equilibrium distribution after $\tau_{\exp}(f)$ time steps [55]. However, determining $\tau_{\exp}(f)$ for every function $f \in \mathcal{H}$ is not feasible. Is there any way to narrow down this search to a select group of functions $\mathcal{H}'$? The answer to this question is "yes". In [16], Fishman states that in practice we only need to consider the set $\mathcal{H}'$ of functions that are

of interest in the study. Furthermore, if the function $f' \in \mathcal{H}'$ has the maximum variance over all functions in $\mathcal{H}'$, we can estimate τ_{exp} for the study to be:

$$\tau_{\text{exp}} = \tau_{\text{exp}}(f'). \tag{3.17}$$

Another name for this estimate of τ_{exp} is the *burntime* of the simulation. Since data coming from time steps before τ_{exp} may not necessarily be from the equilibrium distribution, these datapoints can be discarded to minimize the chance of bias due to non-equilibrium data. However, if τ_{exp} is less than 5% of the total simulation run time, then Sokal [55] argues that the statistical error due to this burntime data is minimal and therefore the data coming from the first τ_{exp} time steps can be included in any estimate.

3.2.3 Estimating τ_{exp} via Warm-up Analysis

The following discussion is based on Section 6.3 in [16]. One method for estimating τ_{exp} is to estimate a finite interval $[0, k]$, called the *warm-up interval for the Markov chain* $\mathcal{X}$, such that $\tau_{\text{exp}} \in [0, k]$. To estimate this interval, one employs n_0 independent replications with t_0 time steps, where the starting state in each replication is the same state $s \in \mathcal{S}$.

Define $\{X_t^r | t \in T\}$ to be the Markov Chain corresponding to the r^{th} replication, where X_t^r is the state at time t in the r^{th} replication. Note that $X_0^r = s$ for every replication r. Given a function $h \in \mathcal{H}$ and $a, b \in T$ such that $a \leq b$, define

$$\bar{h}(r, a, b) = \frac{1}{b - a + 1} \sum_{t=a}^{b} h(X_t^r) \tag{3.18}$$

to be the average of $h(X_t^r)$ in replication r between time steps a and b and

$$\bar{\bar{h}}(a, b) = \frac{1}{n_0} \sum_{r=1}^{n_0} \bar{h}(r, a, b) \tag{3.19}$$

to be the average of $h(X_t^r)$ over all n_0 replications from time steps a to b.

For $1 \leq j \leq t_0$, the quantities $\bar{\bar{h}}(1, j)$ (referred to as the first j column averages) and $\bar{\bar{h}}(t_0 - j + 1, t_0)$ (referred to as the last j column averages) can be used to estimate the warm-up interval $[0, \hat{k}]$. Provided the Markov chain has converged, there should be a point j^* where the trends of the sequences $\bar{\bar{h}}(1, j)$ and $\bar{\bar{h}}(t_0 - j + 1, t_0)$ dissipate for all $j \geq j^*$; this implies that $\hat{k} \leq j^*$. This estimate is a very rough upper bound for $\tau_{\text{exp}}(h)$ [61]. If the function h is the previously discussed function with maximum variance in $\mathcal{H}'$, then j^* can also serve as an upper bound for τ_{exp}.

Although it is likely that the warm-up analysis described in Section 3.2.3 provides an upper bound for τ_{exp}, it is also possible that this estimate can be misleading. Fishman [16, p. 513] states that "starting all replications in the same state leaves the discomforting thought that any ostensible convergence may be due to a local equilibrium,..., and not to the desired global equilibrium. Local stagnation of a process can occur when its equilibrium distribution π is multimodal and its transition matrix $\mathbb{P}$ makes one-step transitions only in a small neighborhood around the current state of the process." In the next section another technique for estimating a warm-up interval will be introduced which addresses these problems.

3.2.4 Estimating τ_{exp} via a Potential Scale Reduction

Suppose now that we have n_0 replications of t_0 time steps, where the i^{th} replication is started in state $s_i \in \mathcal{S}$, and the Markov chain corresponding to that replication is denoted by $\{X_t^i | t \in T\}$. Fishman [16, p. 513] states that if "graphical analysis of $\{X_j^i, 1 \leq j \leq t_0\}$ for $1 \leq i \leq n_0$ reveals a positive integer $k < t_0$ such that all n_0 truncated sample paths have converged to a common region and repeatedly intersect each other, this observation supports the choice of k as an adequate warm-up interval". One numerical method which attempts to quantify the condition "all sample paths have converged to a common region and repeatedly intersect each other" was developed by Gelman and Rubin in [19].

This method is implemented as described in [61]. For each replication, choose starting states $s_1, \ldots, s_{n_0}$ that are "relatively far apart"; this limits the chance that any convergence detected will be from a local equilibrium. Run each replication for t_0 time steps, and for a chosen function $h \in \mathcal{H}'$, define

$$B_{n_0,j} = \frac{j+1}{n_0 - 1} \sum_{r=1}^{n_0} \left(\bar{h}(r,0,j) - \bar{\bar{h}}(0,j) \right)^2 \tag{3.20}$$

to be the *between the replication variance* and

$$W_{n_0,j} = \frac{1}{n_0 j} \sum_{r=1}^{n_0} \sum_{i=0}^{j} \left(h(X_i^r) - \bar{h}(r,0,j) \right)^2 \tag{3.21}$$

to be the *within the replication variance*.

The variance of $h(X_{t_0})$ can be estimated in two ways. The first way to estimate this variance is by:

$$\hat{\text{var}}(h(X_{t_0})) = \frac{t_0}{t_0 + 1} W_{n_0,t_0} + \frac{1}{t_0 + 1} B_{n_0,t_0}. \tag{3.22}$$

This estimate is unbiased under the assumption of stationarity but is an overestimate when the starting states are far apart [18]. The second estimate for this variance is simply W_{n_0,t_0}. Gelman

and Rubin [19] show that as $t_0 \to \infty$,

$$W_{n_0,t_0} \to \text{var}_\pi(h) \tag{3.23}$$

and

$$\hat{\text{var}}(h(X_{t_0})) \to \text{var}_\pi(h). \tag{3.24}$$

Under the assumption that each replication is started in different states that are relatively far apart, [18] shows that the convergence of the Markov chain to its equilibrium distribution can be detected by monitoring the convergence of the sequence $\sqrt{\hat{R}_j}, j \in \{1, \ldots, t_0\}$, where

$$\sqrt{\hat{R}_j} := \sqrt{\frac{\hat{\text{var}}(h(X_j))}{W_{n_0,j}}}. \tag{3.25}$$

The elements of this sequence are referred to as the *estimated potential scale reduction* [19]. For $j \leq t_0$, $\sqrt{\hat{R}_j}$ reduces to:

$$\sqrt{\hat{R}_j} = \sqrt{\frac{j}{j+1} + \frac{1}{j+1}\frac{B_{n_0,j}}{W_{n_0,j}}}. \tag{3.26}$$

As $t_0 \to \infty$, $\sqrt{\hat{R}_{t_0}}$ will converge to 1 [19]. As this happens, replications of the Markov chain become overlapping and satisfy the criteria "all sample paths have converged to a common region and repeatedly intersect each other" [16]. Gelman [18] states that if there exists some $k < t_0$ such that the estimates $\sqrt{\hat{R}_j}$ for all $k \leq j \leq t_0$, are less than 1.1, then the simulation can be thought to have converged for the function h. This k can be considered an upper limit of a warm-up interval for the n_0 replications. The interpretation of this condition is requiring that the estimated standard error between the replications to be less than 10% larger than the estimated standard error within the replications. For the sake of being conservative, the work presented here will require that the estimated standard error between the replications be less than 5% larger than the estimated standard error within the replications; that is, a series of replications will be considered to have converged at time k if $\sqrt{\hat{R}_j} < 1.05 \; \forall \; j \geq k$. This condition is more stringent than the 10% cutoff level for convergence suggested by Gelman in [18]. The reason that 1.05 is selected as the cutoff is so that one can be more confident that the data coming from time steps after the estimated convergence point is indeed coming from the desired equilibrium distribution.

3.2.5 Essentially independent data

Another problem in Markov chain simulations, as mentioned before, is dealing with the correlation between states in the same chain. We expect that the more time steps there are between states, the less correlation there will be between those states. Given a stochastic process $\{f(X_t), t \in T\}$ that is started in its equilibrium distribution, if the correlation between two states is negligible we say that they are *essentially independent*. This concept can be quantified by the *integrated autocorrelation time* of the observable f, called $\tau_{\text{int}}(f)$. $\tau_{\text{int}}(f)$ is defined as follows [55]:

$$\tau_{\text{int}}(f) := \frac{1}{2} \sum_{h=-\infty}^{+\infty} \rho_f(h), \tag{3.27}$$

Where $\rho_f(h)$ is as defined in Equation 3.14. $\tau_{\text{int}}(f)$ can be shown to simplify to:

$$\tau_{\text{int}}(f) = \frac{1}{2} + \sum_{h=1}^{+\infty} \rho_f(h). \tag{3.28}$$

So why is $\tau_{\text{int}}(f)$ a good estimator? Szafron provides the following argument in [61, p. 121]:

Consider the sample mean of f based on $\{f(X_t), t \in T\}$:

$$\bar{f}_n := \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

Then the variance of this estimator is:

$$
\begin{aligned}
\mathrm{var}_\pi(\bar{f}_n) &= E_\pi\left[\frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} f(X_r)f(X_{r+s-r}) - (E_\pi(f))^2\right] \\
&= \frac{\gamma_f(0)}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n}\rho_f(s-r) \\
&= \frac{\gamma_f(0)}{n^2}\sum_{h=-(n-1)}^{n-1}(n-|h|)\rho_f(|h|) \\
&= \frac{\gamma_f(0)}{n}\sum_{h=-(n-1)}^{n-1}\left(1-\frac{|h|}{n}\right)\rho_f(|h|) \\
&= \frac{\gamma_f(0)}{n}\left[2\sum_{h=1}^{n-1}\left(\rho_f(h)-\frac{h\rho_f(h)}{n}\right)+1\right] \\
&= \frac{2\gamma_f(0)}{n}\sum_{h=1}^{n-1}\left(\rho_f(h)-\frac{h\rho_f(h)}{n}+\frac{1}{2(n-1)}\right) \\
&\approx \frac{2\gamma_f(0)}{n}\tau_{\mathrm{int}}(f),\ \text{if } n \gg \tau_{\mathrm{int}}(f),
\end{aligned}
$$

where it is assumed that for $n \gg \tau_{\mathrm{int}}(f)$, $\tau_{\mathrm{int}}(f)$ can be approximated by [55]:

$$
\tau_{\mathrm{int}}(f) \approx \frac{1}{2} + \sum_{1\le h\le n-1}\rho_f(h). \tag{3.29}
$$

The approximation

$$
\mathrm{var}_\pi(\bar{f}_n) \approx \frac{2\gamma_f(0)}{n}\tau_{\mathrm{int}}(f),\ \text{if } n \gg \tau_{\mathrm{int}}(f), \tag{3.30}
$$

implies that the variance of the sample mean is approximately a factor or $2\tau_{\mathrm{int}}(f)$ larger than $\frac{\gamma_f(0)}{n}$, which is the variance of the sample mean computed using independent data. If n values of the observable $f(\mathcal{X})$ are correlated, then there are really $\frac{n}{2\tau_{\mathrm{int}}(f)}$ essentially independent observations [61, p. 121].

Definition 3.2.8. *Given a stochastic process $\{f(X_t), t \in T\}$ with equilibrium distribution π, $f(X_i)$ and $f(X_j)$ are defined to be* essentially independent *if*

$$
|j-i| \ge 2\tau_{\mathrm{int}}(f). \tag{3.31}
$$

Since there is an integrated autocorrelation time for each observable $f \in \mathcal{H}$, it is of interest to know how many time steps must pass before the original states X_i and X_j are independent. This quantity is called the *integrated autocorrelation time for the system* and is defined by [55]:

$$\tau_{\text{int}} := \sup_{f \in \mathcal{H}} \tau_{\text{int}}(f). \tag{3.32}$$

Thus, if we know what function has the highest integrated autocorrelation time, we can use this to estimate the integrated autocorrelation time for the system.

3.2.6 Estimating τ_{int} using Batch Means

$\tau_{\text{int}}(f)$ can be estimated using a procedure known as the "batch means technique" discussed by Fishman in [16]. This technique requires the assumption that all data comes from the equilibrium distribution. Suppose that we have a stochastic process $\{f(X_t), t \in \{0, 1, \ldots, n\}\}$ of length n and an estimate for τ_{exp} of this process, call it k. Consider here only the states from time steps greater than k, namely the set $\{f(X'_t), t \in \{0, 1, \ldots, n-m\}\}$, where $X'_t = X_{t+m}$. For fixed positive integers b and l such that $bl \leq t - m$, consider a set of l 'batches' of this data, each of size b. Here l should be the maximum number of batches of size b one can obtain from the $t - m$ states. For $1 \leq j \leq l$, define the j^{th} batch mean to be:

$$Y_{l,b} := \frac{1}{b} \sum_{i=1}^{b} f(X'_{(l-1)b+i}). \tag{3.33}$$

Also define the average of the batch means to be

$$\bar{F}_{l,b} := \frac{1}{l} \sum_{j=1}^{l} Y_{l,b}, \tag{3.34}$$

and the sample variance of the batch means to be

$$s^2(\bar{F}_{l,b}) := \frac{1}{l-1} \sum_{i=1}^{l} (Y_{l,b} - \bar{F}_{l,b})^2. \tag{3.35}$$

If all of the batch means $Y_{l,b}$ are statistically independent, then b is called an *independent batch size*. The test for independence used here is as follows [16]:

Define the null hypothesis for this test to be $H_0 : \psi_{l,b} = 0$, where

$$\psi_{l,b} := 1 - \frac{\sum_{i=1}^{l-1} (Y_{i,b} - Y_{i+1,b})^2}{2 \sum_{i=1}^{l} (Y_{i,b} - \bar{F}_{l,b})^2}. \tag{3.36}$$

Given a significance level α, the null hypothesis for this test is not rejected when

$$\psi_{l,b} \leq \Phi^{-1}(1 - \alpha) \sqrt{\frac{l-2}{l^2 - 1}}, \tag{3.37}$$

where $\Phi^{-1}(1-\alpha)$ is the $(1-\alpha)$ critical value of the standard normal distribution. It should be noted that in this work we use $\alpha = 0.05$. If the null hypothesis H_0 is not rejected for this α, then the batches are considered to be essentially independent.

Given a sufficiently large sample from the equilibrium distribution, we can make the following approximation (shown in Equation 3.30):

$$\mathrm{var}_\pi(\bar{f}_n) \approx \frac{2\gamma_f(0)}{n}\tau_{\mathrm{int}}(f); \tag{3.38}$$

therefore, an estimate for $\tau_{\mathrm{int}}(f)$ is given by [16]:

$$\hat{\tau}_{\mathrm{int}}(f) = \frac{b}{2}\frac{\hat{\gamma}_f(0)}{s^2(\bar{F}_{l,b})}. \tag{3.39}$$

As the sample size increases, $\hat{\gamma}_f(0) \approx s^2(\bar{F}_{l,b})$ [16]. Thus, $\hat{\tau}_{\mathrm{int}}(f)$ can be approximated simply by $\frac{b}{2}$.

An important question is "how do we know what the best value of b to use in this Equation 3.39 is?". If a batch size b passes the test for independence, then the batch means corresponding to this batch size can be considered essentially independent (at significance level 0.05). However, there is a possibility that an unusually small batch size might pass the test for independence; thus, one could consider batches of that size to be essentially independent. To ensure that such an outlier is never chosen, the for the estimate of τ_{int} in this work is the first batch size b which passes the test for independence such that the batch sizes $b^* > b$ consistently pass the test for independence. This selection procedure reduces the chance of underestimating $\tau_{\mathrm{int}}(f)$.

3.3 Chapter Summary

As one can see, there are many issues that need to be addressed when performing Markov Chain Monte Carlo. One cannot naively assume that all the data coming from a chain is independent, nor can one immediately assume that all the data generated from the chain corresponds to the chain's equilibrium distribution. On the other hand, the methods presented in this chapter for determining τ_{exp} and τ_{int} are only estimates. There is no absolute science to determining exactly when a chain has converged, or exactly how much time must pass before two states are essentially independent. To err on the side of caution, I will tend to be more conservative when addressing these convergence and independence questions.

Recall that the methods described in this chapter have all been used previously by Szafron for the CMC Θ-BFACF Algorithm of the LSP Model [61]. However, as I wrote my own programs to implement these methods (using C and R code), I felt that it was necessary to review these methods again.

Now that the necessary theory of Markov chains has been presented, it is time to discuss specific algorithms designed to generate Markov chains that sample from the distributions of interest relating to the good solvent model presented in Section 2.3.1.

Chapter 4

Algorithms for Generating Random SAPs in a Good Solvent

The following chapter reviews some existing methods for generating random SAPs in the simple cubic lattice where it is assumed that the SAPs are modelling ring polymers in a good solvent. Recall that in this model, two SAPs with the same length are equally likely. The algorithms introduced in this chapter are the pivot algorithm, the BFACF algorithm, and the Θ-BFACF algorithm. The ergodic classes of these algorithms vary; the pivot algorithm samples SAPs with fixed length but variable knot type, while the BFACF algorithm and the Θ-BFACF algorithm sample SAPs or Θ-SAPs of variable length but fixed knot type. These algorithms corresponding to the good solvent case are reviewed here since they are the building blocks for studying the ILSP model.

4.1 Pivot Algorithm

The pivot algorithm [8, 34, 41] is a method for sampling SAPs in the simple cubic lattice, and has been shown to be highly efficient [41]. This algorithm generates a Markov Chain $\{X_t, t \in T\}$ that is ergodic on the set of all SAPs in $\mathcal{P}_n$ and has the target equilibrium distribution [39]

$$\pi_x = \frac{1}{p_n}, \ \ \forall x \ \in \mathcal{P}_n. \tag{4.1}$$

This algorithm attempts to make large scale changes to SAPs, while never changing the number of polygon edges. Although these moves are not accepted very often, when they are, it can radically change a SAP's conformation [39]. These pivot moves alter a selected *segment* of a SAP, where a segment with endpoints $\mathbf{v}_i$ and $\mathbf{v}_j$ is defined to be an ordered sequence of vertices and edges in the SAP that connects vertex $\mathbf{v}_i$ to vertex $\mathbf{v}_j$ (it is obvious that $\mathbf{v}_i$ and $\mathbf{v}_j$ must be part of the original SAP for this to work). It will be assumed that segments specified as having ordered vertices $\mathbf{v}_1, \ldots, \mathbf{v}_n$ also have edges connecting these vertices (*i.e.* $\mathbf{v}_1$ to $\mathbf{v}_2$, etc.).

For SAPs of length n in $\mathbb{Z}^3$, a naive implementation of the pivot algorithm (as described in [41]) has an estimated mean time of $O(n^{0.89})$ per attempted pivot [8]. In 2002, Kennedy [32] found a more efficient way to implement the pivot algorithm; this implementation has an estimated mean time of $O(n^{0.74})$ per attempted pivot. In 2010, Clisby [8] greatly improved on this efficiency in proposing an implementation of the pivot algorithm with an estimated mean time of $O(\log n)$ per attempted pivot. Although the latter two implementations of the pivot algorithm are more efficient, the work presented here uses the naive implementation of the pivot algorithm. This naive implementation was used because it was easier to program, and also because the I-Pivot algorithm (described in Section 5.2) requires performing a calculation at each time step that takes $O(n^2)$ time.

4.1.1 Types of Pivots

There are 13 different types of pivot moves that can be generalized into the three different classes described below:

One type of pivot move is called an *inversion* move, defined as follows. Suppose we have some segment s as defined above with vertices $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{k+1}\}$ and edges connecting these vertices. For each $l = 1, \ldots, k$, define $\vec{d}_l := \mathbf{w}_{l+1} - \mathbf{w}_l$. A new segment s^* with vertices $\{\mathbf{w}_1^*, \ldots, \mathbf{w}_{k+1}^*\}$ can be calculated as follows, where for $b = 1, \ldots, k+1$,

$$\mathbf{w}_b^* = \begin{cases} \mathbf{w}_1, & \text{if } b = 1, \\ \mathbf{w}_{b-1} + \vec{d}_{k+2-b}, & \text{if } 2 \leq b \leq k, \\ \mathbf{w}_{k+1} & \text{if } b = k+1. \end{cases} \tag{4.2}$$

An inversion move is always 'feasible', *i.e.* when the segment s from the SAP is replaced with the segment s^*, the resulting object will still be a polygon (not necessarily self avoiding). An example of an inversion on a segment is shown in Figure 4.1.

Another type of pivot move that can be attempted is called a *reflection* move. These reflections are transformations of a segment through a hyperplane that makes angles of 45 degrees with exactly two of the coordinate hyperplanes [39]. There are 6 types of reflections that can be attempted; however, no more than one reflection will be feasible for any given segment. To see why, suppose we have a segment s with vertices $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{k+1}\}$. Given two dimensions α and β such that $1 \leq \alpha < \beta \leq 3$ (where dimensions 1, 2, and 3 refer to x, y, and z, respectively), the remaining dimension $\delta \neq \alpha, \beta$, and some $m \in \{+1, -1\}$, the reflection $R_{\alpha,\beta,m}$ defined by α, β and m is feasible if and only if [39]

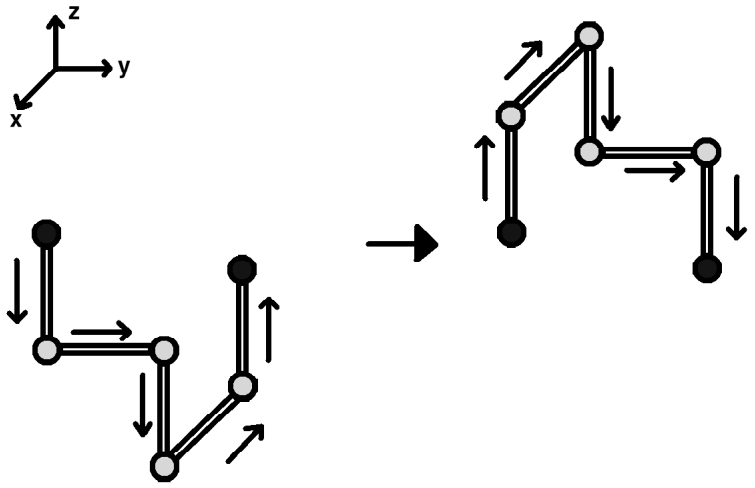

Figure 4.1: An example of an inversion pivot move on a segment.

$$\begin{aligned} \mathbf{w}_{k+1}^{(\alpha)} - \mathbf{w}_1^{(\alpha)} &= m(\mathbf{w}_{k+1}^{(\beta)} - \mathbf{w}_1^{(\beta)}), \quad \text{and} \\ \mathbf{w}_{k+1}^{(\delta)} - \mathbf{w}_1^{(\delta)} &= 0, \end{aligned}$$

where $\mathbf{w}_j^{(d)}$ is the value of the coordinate in the d^{th} dimension of $\mathbf{w}_j$.

Supposing that the reflection $R_{\alpha,\beta,m}$ is feasible, define the new segment that will result from $R_{\alpha,\beta,m}$ to be s^* with vertices $\{\mathbf{w}_1^*, \ldots, \mathbf{w}_{k+1}^*\}$, where for $b = 1, \ldots, k+1$,

$$\mathbf{w}_b^{*(\epsilon)} = \begin{cases} \mathbf{w}_1^{(\epsilon)}, & \text{if } b = 1, \\ \mathbf{w}_{k+2-b}^{(\delta)}, & \text{if } 2 \le b \le k, \text{ and } \epsilon = \delta, \\ \mathbf{w}_1^{(\alpha)} - m\left(\mathbf{w}_{k+2-b}^{(\beta)} - \mathbf{w}_{k+1}^{(\beta)}\right), & \text{if } 2 \le b \le k, \text{ and } \epsilon = \alpha, \\ \mathbf{w}_1^{(\beta)} - m\left(\mathbf{w}_{k+2-b}^{(\alpha)} - \mathbf{w}_{k+1}^{(\alpha)}\right), & \text{if } 2 \le b \le k, \text{ and } \epsilon = \beta, \\ \mathbf{w}_{k+1}^{(\epsilon)} & \text{if } b = k+1. \end{cases} \tag{4.3}$$

An example of one type of reflection on a segment is shown in Figure 4.2.

The last class of pivot moves that can be attempted are called *interchange* moves. There are 6 different types of interchanges, with a maximum of 3 being feasible on any given segment. Suppose we have a segment s with vertices $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{k+1}\}$. Given dimensions α and β, where $1 \le \alpha < \beta \le 3$, and $m \in \{+1, -1\}$, the interchange $N_{\alpha,\beta,m}$ is feasible if and only if [39]

$$\mathbf{w}_{k+1}^{(\alpha)} - \mathbf{w}_1^{(\alpha)} = m(\mathbf{w}_{k+1}^{(\beta)} - \mathbf{w}_1^{(\beta)}).$$

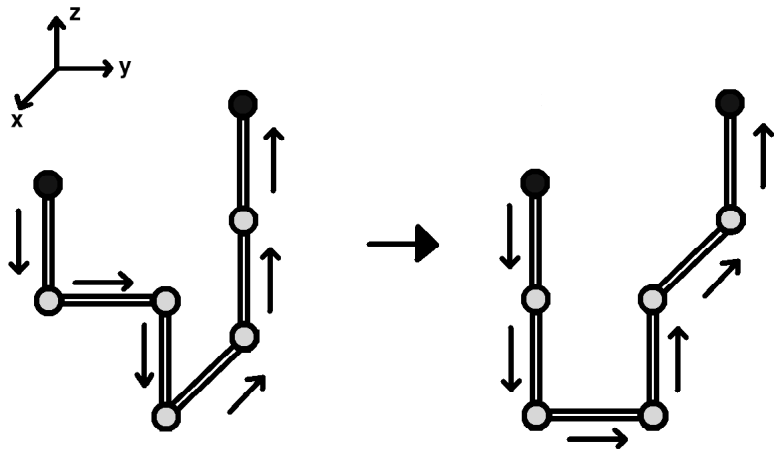

Figure 4.2: An example of the reflection $R_{x,y,-1}$.

Suppose that the interchange $N_{\alpha,\beta,m}$ is feasible. For each $l = 1, \ldots, k$, let $\vec{d_l} := \mathbf{w}_{l+1} - \mathbf{w}_l$. Define the new segment that will result from $N_{\alpha,\beta,m}$ to be s^* with vertices $\{\mathbf{w}_1^*, \ldots, \mathbf{w}_{k+1}^*\}$ where for $b = 1, \ldots, k+1$,

$$\mathbf{w}_b^* = \begin{cases} \mathbf{w}_1, & \text{if } b = 1, \\ \mathbf{w}_{b-1} + \vec{d}_{b-1} & \text{if } 2 \leq b \leq k, \text{ and } |\vec{d}_{b-1}^{(\delta)}| = 1, \\ \mathbf{w}_{b-1} + m * \vec{d}_{b-1}, & \text{if } 2 \leq b \leq k, \text{ and } |\vec{d}_{b-1}^{(\alpha)}| \text{ or } |\vec{d}_{b-1}^{(\beta)}| = 1, \\ \mathbf{w}_{k+1} & \text{if } b = k+1. \end{cases} \tag{4.4}$$

An example of one type of interchange on a segment is shown in Figure 4.3.

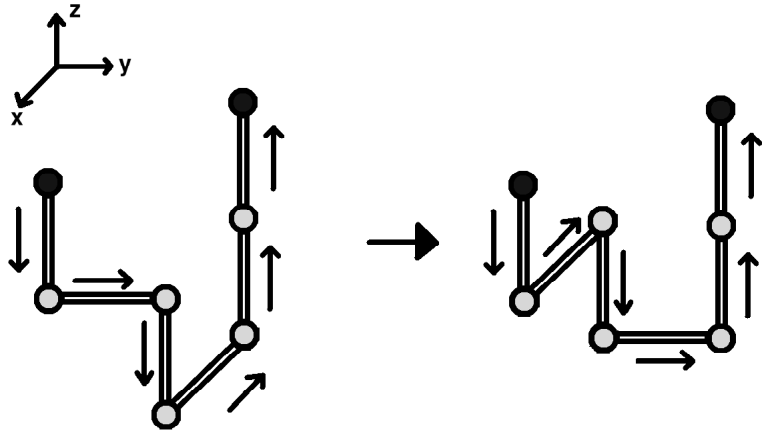

Figure 4.3: An example of the interchange $N_{x,y,-1}$.

It is easy to show that the one step transition probabilities defined by these pivot moves are symmetric; *i.e.* $P_{xy} = P_{yx}$, $\forall\ x, y \in \mathcal{P}_n$. Suppose $\pi_x = \frac{1}{p_n} > 0$, $\forall\ x \in \mathcal{P}_n$. Then $\sum_{x \in \mathcal{P}_n} \pi_x = 1$ and

$\pi_x P_{xy} = \frac{1}{p_n} P_{yx} = \pi_y P_{yx}$, $\forall\, x, y \in \mathcal{P}_n$. Because the pivot algorithm is ergodic on $\mathcal{P}_n$, by Definition 3.1.8 the pivot algorithm is positive recurrent, irreducible and aperiodic. Thus, by Theorems 3.1.10 and 3.1.11, $\pi = \{\pi_x\}_{x \in \mathcal{P}_n}$ is the unique stationary equilibrium distribution of the Markov chain.

4.1.2 Markov Chain using the Pivot Algorithm

A Markov chain $\{X_t, t \in T\}$ using the pivot algorithm is defined as follows:

Start with $t = 0$, and set X_0 equal to some starting SAP ω_0 of length n.

Suppose that $(\mathbf{v}_1, \ldots, \mathbf{v}_n)$ are the ordered vertices of X_t. Choose an integer uniformly at random from the set $\{1, \ldots, n\}$; call this integer i^*. Choose another integer uniformly at random from the set $\{1, \ldots, n\} \setminus \{i^*\}$; call this integer j^*. Choose the shortest segment of edges in X_t connecting $\mathbf{v}_{i^*}$ to $\mathbf{v}_{j^*}$; if both segments are of equal length, choose either one with probability 0.5. Call this chosen segment $s_{i^*j^*}$. Suppose that $s_{i^*j^*}$ has the k ordered edges $(e_1^*, \ldots, e_k^*)$ and $k+1$ ordered vertices $(\mathbf{v}_{i^*} =: \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{k+1} := \mathbf{v}_{j^*})$ connecting $\mathbf{v}_{i^*}$ to $\mathbf{v}_{j^*}$.

Determine which pivots are feasible for the chosen segment, and uniformly at random choose one of these feasible pivots to attempt. Define the segment that results from the pivot on $s_{i^*j^*}$ to be $s^*_{i^*j^*}$. Define X_{prop} to be X_t with $s_{i^*j^*}$ replaced by the newly created segment $s^*_{i^*j^*}$. If X_{prop} is a SAP, then set $X_{t+1} = X_{\text{prop}}$; otherwise reject the pivot and set $X_{t+1} = X_t$.

Increment t by 1 and repeat the following procedure starting with selecting a new segment.

4.2 BFACF Algorithm

The *BFACF algorithm* [4, 11, 12] generates a Markov chain on the set of all self-avoiding polygons in $\mathbb{Z}^3$ with *a priori* chosen knot type K. This chain has the target equilibrium distribution defined by Equation 2.34 with $w(n) = n^q$, where q is a positive integer [61]; *i.e.*

$$\pi_\omega(q, z) = \frac{|\omega|^q z^{|\omega|}}{Q_K(z, w)}; \forall\, \omega \in \mathcal{P}(K). \tag{4.5}$$

It has been proven [70] that the BFACF algorithm is ergodic on the set of all SAPs with knot type K. This implies that if one starts with a SAP ω with knot type K, then using the BFACF

algorithm, it is possible to obtain any other SAP ω' with the same knot type. It also implies that the BFACF algorithm will never change a SAP's knot type. The parameter z in the BFACF algorithm is called the *fugacity* of the chain. Increasing z will yield larger polygons on average at equilibrium. The purpose of choosing $w(n) = n^q$ in the equilibrium distribution of the BFACF algorithm is discussed further below.

The definition of the BFACF algorithm is as follows [60]:

Given a knot type K, select an integer $q > 0$ and z such that $0 < z < z_c(K)$, where

$$-\log z_c(K) := \limsup_{n\to\infty} \log\left((2n)^q p_{2n}(K)\right)^{\frac{1}{2n}} = \limsup_{n\to\infty} \frac{1}{2n}\log p_{2n}(K) = \kappa_K, \tag{4.6}$$

and κ_K is as defined by Equation 1.3.15. Note that $z_c(K) = e^{-\kappa_K}$, called the *critical z-value*, is the radius of convergence of $Q_K(z,w)$. Next, choose an initial SAP $\omega^{[0]} \in \mathcal{P}(K)$. Set $t = 0$, $X_0 = \omega^{[0]}$, and select one of the vertices of $\omega^{[0]}$ to be denoted $\omega_0^{[0]}$. Now select one of the two vertices of $\omega^{[0]}$ connected by an edge to $\omega_0^{[0]}$ and denote this vertex to be $\omega_1^{[0]}$. $\omega_0^{[0]}$ is referred to as the first vertex of $\omega^{[0]}$ and $\omega_1^{[0]}$ is referred to as the second vertex of $\omega^{[0]}$. The edge connecting $\omega_0^{[0]}$ to $\omega_1^{[0]}$ will be referred to as the 1^{st} edge of $\omega^{[0]}$. This numbering of vertices imposes an ordering on the SAP; number the remaining edges and vertices using this ordering. Choose a set of one-step transition probabilities $P_{\omega\omega'}$ such that these probabilities satisfy $\pi_\omega(q,z)P_{\omega\omega'} = P_{\omega'\omega}\pi_{\omega'}(q,z)$ and such that $\lim_{n\to\infty} P^{(n)}_{\omega\omega'} = \pi_{\omega'}(q,z)$.

Starting with X_t, $t = 0$, the Markov chain proceeds as follows:

Number the vertices and edges of X_t as described above and select an integer i of X_t uniformly at random between 1 and $|X_t| = n$; this corresponds to selecting the i^{th} edge of X_t. Consider the 4 graph embeddings W_1, W_2, W_3, and W_4 in $\mathbb{Z}^3$ that result from moving this i^{th} edge one lattice unit in each of the 4 unit directions perpendicular to that edge, and then adding the necessary edges to join the newly shifted edge to the SAP, removing any double edges that might result. The process of "shifting of an edge" and rejoining it to the SAP is called a *BFACF move*. A BFACF move can only add two edges (called a *p(+2) move*), remove two edges (called a *p(-2) move*), or leave the number of edges unchanged (called a *p(0) move*); examples of these moves are shown in Figure 4.4.

For each $i = 1, 2, 3, 4$, define the probability of proposing the embedding W_i to be:

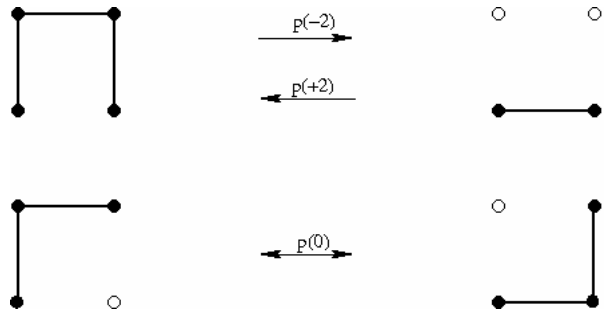

Figure 4.4: Types of BFACF moves (from [61], with permission from the author)

$$\Pr(W_i) := \begin{cases} \frac{(n+2)^{q-1}z^2}{n^{q-1}+3(n+2)^{q-1}z^2} =: p_n(+2), & \text{if } |W_i| - |X_t| = 2, \\ \frac{(n-2)^{q-1}}{(n-2)^{q-1}+3n^{q-1}z^2} =: p_n(-2), & \text{if } |W_i| - |X_t| = -2, \\ \frac{p_n(+2)+p_n(-2)}{2} := p_n(0), & \text{if } |W_i| - |X_t| = 0. \end{cases} \tag{4.7}$$

Define $1 - \sum_{i=1}^{4} \Pr(W_i) =: \Pr(X_t)$ to be the probability of doing nothing and setting $X_{t+1} = X_t$. Choose one of the 5 embeddings W_1, W_2, W_3, W_4, X_t according to their respective probabilities, and denote this chosen embedding to be W. If W is not a SAP, set $X_{t+1} = X_t$, otherwise, set $X_{t+1} = W$. Increment t by 1 and repeat the above procedure.

A detailed discussion on why the transition probabilities defined by the above procedure are a valid choice for the target equilibrium distribution can be found in [60].

When programming this algorithm, if one stores the coordinates of all the vertices occupied by the current SAP in a hash table, then it is possible to check whether a particular vertex is occupied in constant time. Also, if the computer running the algorithm has enough memory to be able to have a pointer corresponding to each vertex in the SAP, then, at each time step, one can look up the edge randomly selected by the algorithm in constant time as well. If both of these features are implemented into the algorithm, then the BFACF algorithm runs in $O(1)$ time.

Now we discuss the bounds on the choice of q; recall that q is a positive integer. In a Markov Chain generated by the BFACF algorithm with parameters q, $z = e^{\beta} < z_c(K)$, and a knot type K, the average length of a SAP in the chain at equilibrium, denoted by $E_{z,w,K}[n]$, is:

$$
\begin{aligned}
E_{z,w,K}[n] &= \sum_{n=1}^{\infty} \frac{n p_n(K) n^q z^n}{Q_K(z,w)} \\
&= \frac{1}{Q_K(z,w)} \sum_{n=1}^{\infty} p_n(K) n^{q+1} z^n \\
&= \frac{1}{Q_K(e^\beta,w)} \sum_{n=1}^{\infty} p_n(K) n^{q+1} e^{\beta n} \\
&= \frac{\partial}{\partial \beta} \log Q_K(e^\beta, w).
\end{aligned}
$$

Thus, if $Q_K(z,w)$ diverges when $z = z_c(K)$, we expect that $E_{z,w,K}[n]$ will also diverge. Since the goal is to sample large polygons, it is preferable that $Q_K(z,w)$ diverges when $z = z_c(K)$.

How can we tell if $Q_K(z,w)$ (and consequently, $E_{z,w,K}[n]$) will diverge when $z = z_c(K)$? As z approaches $z_c(K)$ from below, it is a standard assumption that $Q_K(z,w)$ scales like $\left(1 - \frac{z}{z_c(K)}\right)^t$ [61, Section 4.5]. Therefore, if $t < 0$, $Q_K(z,w)$ will diverge as $z \to z_c(K)$. Assuming that the scaling form for $p_{2n}(K)$ given in Equation 1.12 is valid, Szafron [61] showed that

$$
t = -\alpha_K - q + 2, \tag{4.8}
$$

and also that if $t < 0$ and $w = n^q$, then up to first order

$$
E_{z,w,K}[n] \approx (\alpha_K + q - 2) \left[\frac{z/z_c(K)}{1 - z/z_c(K)} \right], \tag{4.9}
$$

as $z \to z_c(K)$.

To achieve $t < 0$, we want to select q that is greater than $2 - \alpha_K$. In the case of $K = \phi$, α_ϕ has been estimated to be 0.27 ± 0.02 [46]. Therefore, q must be 2 or larger for $Q_\phi(z,w)$ to diverge at $z = z_c(\phi)$. From my direct experience, it is incredibly difficult to sample large polygons when $q = 1$ and $K = \phi$, and increasing q to 2 resolves this problem. This problem is mentioned here because it appears that similar issues arise in some cases of the I-Θ-BFACF algorithm (discussed in Sections 5.5 and 8.5).

In addition to resolving the problem described above, increasing q also yields larger average polygon lengths at equilibrium. One can see this by referring to Equation 4.7 and noting that $p_n(+2)$ increases and $p_n(-2)$ decreases as q increases. Thus, increasing q makes it more likely to perform BFACF moves which add two edges and less likely to perform BFACF moves that remove two edges (for a given set of proposed embeddings). However, there is an upper bound on q in the form of [60]:

$$q \leq 2\frac{\log z_c(K)}{\log \frac{2}{3}} + 1. \tag{4.10}$$

This restriction is required to ensure that the sum of the probabilities of the four embeddings considered at each time step (*i.e.* $\sum_{i=1}^{4} \Pr(W_i)$) will always be less than or equal to 1. As $z_c(\phi)$ has been estimated to be approximately 0.2135 [46], this yields an upper bound of 8 for q (when $K = \phi$). Next, we review how to apply the BFACF algorithm to Θ-SAPs.

4.3 Θ-BFACF Algorithm

Developed and used by Szafron in [60, 61] and Szafron and Soteros in [62, 63], the Θ-BFACF algorithm is a modification of the BFACF algorithm with two main differences; the first difference is that the algorithm is defined only for Θ-SAPs, and the second is that a BFACF move is not allowed to alter one of the edges of the Θ-structure. This chain has the target equilibrium distribution defined by Equation 2.33 with $w(n) = (n-6)n^{q-1}$, where q is a positive integer [61], *i.e.*

$$\pi_{\omega}^{\Theta}(q, z) = \frac{(|\omega| - 6)|\omega|^{q-1} z^{|\omega|}}{Q_K^{\Theta}(z, w)}, \ \forall\, \omega \in \mathcal{P}^{\Theta}(K). \tag{4.11}$$

Given a starting Θ-SAP with knot type K, the Θ-BFACF algorithm generates a Markov chain $\{X_t, t \in T\}$ that is ergodic on the set $\mathcal{P}^{\Theta}(K)$ [60]. The fact that this algorithm preserves the Θ-structure of SAPs throughout is very useful. When combined with the fact that it is ergodic on $\mathcal{P}^{\Theta}(K)$, one can use the Θ-BFACF algorithm to generate essentially independent samples of Θ-SAPs with a particular knot type and connection class. Similarly to the BFACF algorithm, increasing q and z will yield larger polygons on average [61].

The definition of the algorithm is as follows:

Given a knot type K, select an integer $q > 0$ and a fugacity z such that $0 < z < z_c^{\Theta}(K)$, where

$$-\log z_c^{\Theta}(K) := \limsup_{n\to\infty} \log\left((2n)^q p_{2n}^{\Theta}(K)\right)^{\frac{1}{2n}} = \limsup_{n\to\infty} \frac{1}{2n} \log p_{2n}^{\Theta}(K) = \kappa_K^{\Theta} = \kappa_K, \tag{4.12}$$

κ_K is as defined by Equation 1.3.15, and κ_K^{Θ} is as defined by Equation 2.4. Note that $z_c^{\Theta}(K) = e^{-\kappa_K} = z_c(K)$ is the critical z-value for the Θ-BFACF algorithm with knot type K. Moreover,

$z_c^{\Theta}(K)$ is the radius of convergence of $Q_K^{\Theta}(z,w)$. In the case where $K=\phi$, it has been proven [61] that $\kappa_\phi^{\Theta}=\kappa_\phi$, so $z_c^{\Theta}(\phi)=e^{-\kappa_\phi}=z_c(\phi)$.

Next, choose an initial SAP $\omega^{[0]}\in\mathcal{P}^{\Theta}(K)$. Set $t=0$, $X_0=\omega^{[0]}$, and denote the vertex $(0,0,0)$ in $\omega^{[0]}$ (*i.e.* the middle vertex of the top strand of the Θ-structure) to be $\omega_0^{[0]}$. Now select one of the two vertices of $\omega^{[0]}$ connected by an edge to $\omega_0^{[0]}$ and denote this vertex to be $\omega_1^{[0]}$. Number the rest of the vertices in $\omega^{[0]}$ according to the orientation imposed by $\omega_0^{[0]}$ and $\omega_1^{[0]}$. Define the edge connecting $\omega_0^{[0]}$ to $\omega_1^{[0]}$ to be the first edge of $\omega^{[0]}$, the edge connecting $\omega_1^{[0]}$ to $\omega_2^{[0]}$ to be the second edge of $\omega^{[0]}$, and so on.

Choose a set of one-step transition probabilities $P_{\omega\omega'}$ such that these probabilities satisfy $\pi_\omega^{\Theta}(q,z)P_{\omega\omega'}=P_{\omega'\omega}\pi_{\omega'}^{\Theta}(q,z)$ and such that $\lim_{n\to\infty}P_{\omega\omega'}^{(n)}=\pi_{\omega'}^{\Theta}(q,z)$.

Similarly to the BFACF algorithm, an increase in z or q will lead to an increase in the average length of a polygon from the chain.

Starting with X_t, $t=0$, the Markov chain proceeds as follows:

Number the vertices and edges of X_t as described above, uniformly at random select an edge of X_t which is not part of the Θ-structure; call this edge e^*. Similarly to the BFACF algorithm, consider the 4 graph embeddings W_1,W_2,W_3, and W_4 in $\mathbb{Z}^3$ that result from moving e^* one lattice unit in each of the 4 unit directions perpendicular to e^*, and then adding the necessary edges to join the newly shifted edge to the Θ-SAP, removing any double edges that might result. For each $i=1,2,3,4$, define the probability of proposing the embedding W_i to be:

$$\Pr(W_i):=\begin{cases}\frac{(n+2)^{q-1}z^2}{n^{q-1}+3(n+2)^{q-1}z^2}=:p_n(+2), & \text{if } |W_i|-|X_t|=2,\\ \frac{(n-2)^{q-1}}{(n-2)^{q-1}+3n^{q-1}z^2}=:p_n(-2), & \text{if } |W_i|-|X_t|=-2,\\ \frac{p_n(+2)+p_n(-2)}{2}=:p_n(0), & \text{if } |W_i|-|X_t|=0.\end{cases}\tag{4.13}$$

Define $1-\sum_{i=1}^4\Pr(W_i)=\Pr(X_t)$ to be the probability of doing nothing and setting $X_{t+1}=X_t$. Choose one of the 5 embeddings W_1,W_2,W_3,W_4,X_t according to their respective probabilities, and denote this chosen embedding to be W. If W is not a Θ-SAP (*i.e.* it is not self-avoiding or the Θ-structure in X_t was altered), set $X_{t+1}=X_t$, otherwise, set $X_{t+1}=W$. Increment t by 1 and repeat the above procedure.

For a detailed argument explaining the ergodicity of the Θ-BFACF algorithm as defined here, see [60, Section 5.4]. Similarly to the BFACF algorithm, it is possible that the Θ-BFACF algorithm

can run in $O(1)$ time.

4.4 Chapter Summary

This chapter described some algorithms (the Pivot, BFACF and Θ-BFACF algorithms) that are designed to generate Markov chains with equilibrium distributions of interest related to the good solvent model. Using the pivot algorithm, one can sample SAPs of a fixed length and variable knot type (*i.e.* $\mathcal{P}_n$). Using the BFACF algorithm or the Θ-BFACF algorithm one can sample SAPs or Θ-SAPs of a fixed knot type with variable length (*i.e.* $\mathcal{P}(K)$ or $\mathcal{P}^{\Theta}(K)$). For different problems one needs to use different algorithms; for example, when studying the probability of knotting, it is necessary to have an algorithm that can switch between knot types (*e.g.* the pivot algorithm). On the other hand, if one is interested in studying strand passage action on SAPs with a particular knot type, then the Θ-BFACF algorithm is more useful. The next chapter will discuss how to modify these algorithms to sample from distributions according to different solvent conditions.

Chapter 5

Algorithms for Generating Random SAPs in Varying Solvents

The main questions pertaining to this thesis (*i.e.* Problems 1 and 2) are related to finding a good way to model ring polymers and how knot transition probabilities vary with different salt concentrations in solution. To address Problem 1 it is necessary to incorporate interactions that occur with polymers in a salt solution into the model. This was done previously in [64], and in this chapter we review their approach. To address Problem 2, the Θ-BFACF algorithm needs to be modified to incorporate this salt model. The following chapter presents a new algorithm, called the *Interacting Θ-BFACF Algorithm*, which uses Metropolis sampling to sample Θ-SAPs of a particular knot type K based on *a priori* chosen solvent conditions. This chapter also presents some theoretical results pertaining to this new model.

5.1 Metropolis Sampling based on the Energy of a SAP

The pivot, BFACF, and Θ-BFACF algorithms assumed that all conformations of the same length (*i.e.* all SAPs or Θ-SAPs with the same number of edges) are equally likely. Because DNA is negatively charged and interacts with the solution in which it exists, it is useful to have this interaction incorporated in the model. Using Metropolis sampling based on the energy of a SAP is an ideal way to introduce this interaction into the model.

5.1.1 Metropolis Sampling Definition

The following discussion is based on [16, Section 5.14]. Suppose we have a Markov chain that is ergodic on a set $\mathcal{S}$ with one-step transition probabilities defined by $R := \{R_{xy} | x, y \in \mathcal{S}\}$. Suppose the target equilibrium distribution is $\pi = \{\pi_x > 0\}_{x \in \mathcal{S}}$. It is possible to sample from this equilibrium distribution using the Markov chain $\{X_t, t \in T\}$ on $\mathcal{S}$ with one step transition probabilities

$\{P_{xy}|x, y \in \mathcal{S}\}$ defined by [16]:

$$P_{xy} := \begin{cases} \alpha_{xy} R_{xy}, & \text{if } x \neq y, \\ 1 - \sum_{j \in \mathcal{S}, j \neq x} P_{xj}, & \text{if } x = y, \end{cases} \tag{5.1}$$

where $R_{yx} = 0$ if $R_{xy} = 0$, and

$$\alpha_{xy} := \begin{cases} 1, & \text{if } R_{xy} = 0, \\ \dfrac{t_{xy}}{1 + \pi_x R_{xy} / \pi_y R_{yx}}, & \text{if } R_{xy} > 0, \end{cases} \tag{5.2}$$

and $\{t_{xy} = t_{yx}\}$ is chosen to ensure that $0 < \alpha_{xy} \leq 1, \ \forall \ x, y \in \mathcal{S}$ for which R_{xy} and $R_{yx} > 0$. Using a Markov Chain with the one-step transition probabilities defined by P_{xy} in Equations 5.1 and 5.2 is sometimes referred to as the Metropolis-Hastings method [26].

Theorem 5.1.1 ([16])**.** *A Markov chain with the one-step transition probabilities $P_{xy} = \alpha_{xy} R_{xy}$ as defined by Equations 5.1 and 5.2 is reversible. It is also irreducible and aperiodic provided that the Markov chain with one step transition probabilities defined by R is irreducible and aperiodic; moreover,*

$$\sum_{x \in \mathcal{S}} \pi_x P_{xy} = \pi_y, \ \forall \ y \in \mathcal{S}. \tag{5.3}$$

Thus, if the Markov chain defined by the one step transition probabilities $R := \{R_{xy}|x, y \in \mathcal{S}\}$ is irreducible and aperiodic, then by Definitions 3.1.7 and 3.1.8, the Markov chain with one-step transition probabilities $\mathbf{P} = \{P_{xy}|x, y \in \mathcal{S}\}$ as defined by Equations 5.1 and 5.2 is ergodic with unique equilibrium distribution π.

For a given transition matrix R, one possible choice for t_{xy} is [16]:

$$t_{xy} = \begin{cases} 1 + \dfrac{\pi_x R_{xy}}{\pi_y R_{yx}}, & \text{if } \pi_x R_{xy} \leq \pi_y R_{yx}, \\ 1 + \dfrac{\pi_y R_{yx}}{\pi_x R_{xy}}, & \text{if } \pi_x R_{xy} > \pi_y R_{yx}, \end{cases} \tag{5.4}$$

in which case

$$\alpha_{xy} = \min\left(\frac{\pi_y R_{yx}}{\pi_x R_{xy}}, 1\right). \tag{5.5}$$

As recommended by Fishman [16], the choice of t_{xy} and corresponding choice of α_{xy} as defined in Equations 5.4 and 5.5 will be used in this work.

5.2 The Interacting Pivot Algorithm

Assuming that the starting state is a SAP of length n and the energy parameters are defined by $\mathcal{E} = \{A, \zeta, v, T\}$, using the pivot algorithm with Metropolis sampling based on SAP energy we can sample from the target equilibrium distribution (refer to Equation 2.35)

$$\pi_\omega = \frac{e^{\frac{-U_\mathcal{E}(\omega)}{k_B T}}}{Z_n(\mathcal{E})} \tag{5.6}$$

over all SAPs $\omega \in \mathcal{P}_n$. Because the transition probabilities of the pivot algorithm are symmetric (*i.e.* $R_{xy} = R_{yx}$ for all $x, y \in \mathcal{P}_n$), the Metropolis sampling acceptance rate α_{xy} is:

$$\alpha_{xy} := \begin{cases} 1, & \text{if } R_{xy} = 0, \\ \min\left(e^{\frac{U_\mathcal{E}(x)}{k_B T} - \frac{U_\mathcal{E}(y)}{k_B T}}, 1\right), & \text{if } R_{xy} > 0. \end{cases} \tag{5.7}$$

To implement a simulation of the pivot algorithm with Metropolis sampling (referred to here as the *Interacting Pivot Algorithm* or *I-Pivot Algorithm*), supposing the current state in the chain is $X_t = \omega$, use the pivot algorithm procedure outlined in Section 4.1 to propose a new SAP ω'. This new SAP is accepted as state X_{t+1} with probability $\alpha_{\omega\omega'}$.

5.3 The Interacting Θ-BFACF Algorithm

This new algorithm, referred to here as the *Interacting Θ-BFACF Algorithm* or *I-Θ-BFACF Algorithm* for short, was devised in order to answer questions related to how knot transition probabilities of ring polymers modelled by Θ-SAPs with a fixed knot type vary with different salt concentrations. Much work has been done in studying Θ-SAPs [60, 61, 62, 63]; however, it is assumed in these sources that the Θ-SAPs are in a good solvent where the effect of any interactions that occur is negligible.

Assuming that the starting state of the chain is a Θ-SAP with knot type K, the fugacity of the chain is z, q is a positive integer, and the energy parameters being used are defined by $\mathcal{E} = \{A, \zeta, v, T\}$, using the I-Θ-BFACF algorithm the target equilibrium distribution is given by Equation 2.36 with $w(n) = (n-6)n^{q-1}$, *i.e.*

$$\pi_\omega(q, z, \mathcal{E}) := \frac{e^{\frac{-U_\mathcal{E}(\omega)}{k_B T}}(|\omega| - 6)|\omega|^{q-1} z^{|\omega|}}{Q^\Theta_{K,\mathcal{E}}(z, w)}. \tag{5.8}$$

The above equation contains terms related to the energy of a SAP as well as its length. In this distribution, SAPs with equal length and equal energy are equally likely.

5.3.1 Radius of Convergence of $\mathbf{Q}^{\Theta}_{\mathbf{K},\mathcal{E}}(\mathbf{z},\mathbf{w})$

A natural question one might ask is "what is the radius of convergence of $Q^{\Theta}_{K,\mathcal{E}}(z,w)$?". The answer to this question is that this radius of convergence depends on the energy parameters being used. However, a first step that can be taken is in the following theorem:

Theorem 5.3.1. *The radius of convergence of $Q^{\Theta}_{K,\mathcal{E}}(z,w)$ is positive for any choice of $\mathcal{E} = \{A, \zeta, T, v\}$, where $A \geq 0$, $\zeta > 0$, $T > 0$ and $v \leq 0$.*

Proof. Recall from Equation 2.28 that $U_{\zeta,A,T,v}(\omega) = C(\omega)k_BTv + D_{A,\zeta}(\omega)$, where $C(\omega) \geq 0$ and $D_{A,\zeta}(\omega) \geq 0$. Thus,

$$\begin{aligned} Q^{\Theta}_{\mathcal{E},K}(z,w) &= \sum_{n=1}^{\infty}\left[\sum_{\omega'\in\mathcal{P}^{\Theta}_{2n}(K)} e^{\frac{-U_{\mathcal{E}}(\omega')}{k_BT}}(2n-6)(2n)^{q-1}z^{2n}\right] \\ &= \sum_{n=1}^{\infty}\left[\sum_{\omega'\in\mathcal{P}^{\Theta}_{2n}(K)} e^{-C(\omega')v-\frac{D(\omega')}{k_BT}}(2n-6)(2n)^{q-1}z^{2n}\right]. \end{aligned}$$

Because a SAP with length $2n$ can have no more than $6\times 2n$ contacts, the following inequalities hold:

$$\begin{aligned} &\sum_{n=1}^{\infty}\left[\sum_{\omega'\in\mathcal{P}^{\Theta}_{2n}(K)} e^{-C(\omega')v-\frac{D(\omega')}{k_BT}}(2n-6)(2n)^{q-1}z^{2n}\right] \\ \leq\ &\sum_{n=1}^{\infty}\left[\sum_{\omega'\in\mathcal{P}^{\Theta}_{2n}(K)} e^{-C(\omega')v}(2n-6)(2n)^{q-1}z^{2n}\right] \\ \leq\ &\sum_{n=1}^{\infty}\left[\sum_{\omega'\in\mathcal{P}^{\Theta}_{2n}(K)} e^{-6(2n)v}(2n-6)(2n)^{q-1}z^{2n}\right] \\ =\ &\sum_{n=1}^{\infty} p^{\Theta}_{2n}(K)e^{-6(2n)v}(2n-6)(2n)^{q-1}z^{2n} \\ =\ &\sum_{n=0}^{\infty} l^{\Theta}_{n}(K)z^{n} =: E^{\Theta}_{K}(z), \end{aligned}$$

where

$$l_n^{\Theta}(K) = \begin{cases} 0, & \text{if } n \text{ is odd,} \\ p_n^{\Theta}(K) e^{-12nv}(n-6)(n)^{q-1}, & \text{if } n \text{ is even.} \end{cases} \tag{5.9}$$

Define $z_c(*)$ to be the radius of convergence of $E_K^{\Theta}(z)$. Then

$$\frac{1}{z_c(*)} = \limsup_{n\to\infty} |l_n^{\Theta}(K)|^{\frac{1}{n}} = \limsup_{n\to\infty} \left(p_{2n}^{\Theta}(K) e^{-12nv}(2n-6)(2n)^{q-1}\right)^{\frac{1}{2n}}.$$

This implies that

$$\begin{aligned} -\log z_c(*) &= \limsup_{n\to\infty} \left(\frac{1}{2n} \log p_{2n}^{\Theta}(K) - 6v \right) \\ &= -6v + \kappa_K. \end{aligned}$$

Therefore,

$$z_c(*) = e^{6v-\kappa_K} > 0.$$

Now, $Q_{\mathcal{E},K}^{\Theta}(z,w)$ can be written as a power series as follows:

$$\begin{aligned} Q_{\mathcal{E},K}^{\Theta}(z,w) &= \sum_{n=1}^{\infty} \left[\sum_{\omega' \in \mathcal{P}_{2n}^{\Theta}(K)} e^{-C(\omega')v - \frac{D(\omega')}{k_B T}} (2n-6)(2n)^{q-1} z^{2n} \right] \\ &= \sum_{n=1}^{\infty} a_n^{\Theta,\mathcal{E}}(K) z^n, \end{aligned}$$

where

$$a_n^{\Theta,\mathcal{E}}(K) = \begin{cases} 0, & \text{if } n \text{ is odd,} \\ \sum_{\omega' \in \mathcal{P}_n^{\Theta}(K)} e^{-C(\omega')v - \frac{D(\omega')}{k_B T}} (n-6) n^{q-1}, & \text{if } n \text{ is even.} \end{cases} \tag{5.10}$$

Define $z_c^{\Theta,\mathcal{E}}(K)$ to be the radius of convergence of $Q_{\mathcal{E},K}^{\Theta}(z,w)$. Since $0 \le a_n^{\Theta,\mathcal{E}}(K) \le l_n^{\Theta}(K)$ for all natural numbers n, this implies that

$$\limsup_{n\to\infty} |a_n^{\Theta,\mathcal{E}}(K)|^{\frac{1}{n}} \le \limsup_{n\to\infty} |l_n^{\Theta}(K)|^{\frac{1}{n}},$$

which implies that $z_c^{\Theta,\mathcal{E}}(K) > z_c(*) > 0$. Thus, the radius of convergence of $Q_{K,\mathcal{E}}^{\Theta}(z,w)$ is positive for any proper choice of energy parameters.

□

The consequence of Theorem 5.3.1 is that for any appropriate choice of energy parameters, there will always have some choice of $z > 0$ that will make Equation 5.8 a valid probability distribution (*i.e.* $Q^{\Theta}_{K,\mathcal{E}}(z,w)$ will be finite).

5.3.2 Determining the Acceptance Probability $\alpha_{\mathbf{xy}}$

In order to completely define the I-Θ-BFACF algorithm, the acceptance probability $\alpha_{\omega\omega'}$ needs to be determined for all ω and ω' in the state space. Assume that the Markov chain starts with a Θ-SAP with knot type K, the energy parameters are defined by $\mathcal{E}$, the fugacity of the chain is $0 < z < z_c^{\Theta,\mathcal{E}}(K)$, and q is a positive integer. The purpose of z and q is analogous to its purpose in the Θ-BFACF algorithm. Suppose that $\omega, \omega' \in \mathcal{P}^{\Theta}(K)$, $|\omega| = n$, ω' can be obtained from ω (and vice-versa) in one Θ-BFACF move with one step transition probability $R_{\omega\omega'}$, and $|\omega'| - |\omega| = 2$ (*i.e.* the Θ-BFACF move which takes ω to ω' is a $p(+2)$ move). The acceptance probability for this move, as defined by Equation 5.2, is:

$$\alpha_{\omega\omega'} = \min\left(\frac{\pi_{\omega'}(q,z,\mathcal{E})R_{\omega'\omega}}{\pi_{\omega}(q,z,\mathcal{E})R_{\omega\omega'}}, 1\right). \tag{5.11}$$

If $\dfrac{\pi_{\omega'}(q,z,\mathcal{E})R_{\omega'\omega}}{\pi_{\omega}(q,z,\mathcal{E})R_{\omega\omega'}} < 1$, then

$$\begin{aligned}
\alpha_{\omega\omega'} &= \frac{\dfrac{e^{\frac{-U_{\mathcal{E}}(\omega')}{k_BT}}((n+2)-6)(n+2)^{q-1}z^{n+2}}{Q_{\mathcal{E},K}(z,w)} \times \dfrac{1}{(n+2)-6}\left(\dfrac{n^{q-1}}{n^{q-1}+3(n+2)^{q-1}z^2}\right)}{\dfrac{e^{\frac{-U_{\mathcal{E}}(\omega)}{k_BT}}(n-6)n^{q-1}z^n}{Q_{\mathcal{E},K}(z,w)} \times \dfrac{1}{n-6}\left(\dfrac{(n+2)^{q-1}z^2}{n^{q-1}+3(n+2)^{q-1}z^2}\right)} \\
&= \exp\left(\frac{U_{\mathcal{E}}(\omega)}{k_BT} - \frac{U_{\mathcal{E}}(\omega')}{k_BT}\right).
\end{aligned}$$

Supposing now that $\omega, \omega' \in \mathcal{P}^{\Theta}(K)$, $|\omega| = n$, $|\omega'| = n-2$, ω' can be obtained from ω via one $p(-2)$ Θ-BFACF move and $\dfrac{\pi_{\omega'}(q,z,\mathcal{E})R_{\omega'\omega}}{\pi_{\omega}(q,z,\mathcal{E})R_{\omega\omega'}} < 1$, then

$$\begin{aligned}
\alpha_{\omega\omega'} &= \frac{\dfrac{e^{\frac{-U_{\mathcal{E}}(\omega')}{k_BT}}((n-2)-6)(n-2)^{q-1}z^{n-2}}{Q_{\mathcal{E},K}(z,w)} \times \dfrac{1}{(n-2)-6}\left(\dfrac{n^{q-1}z^2}{(n-2)^{q-1}+3n^{q-1}z^2}\right)}{\dfrac{e^{\frac{-U_{\mathcal{E}}(\omega)}{k_BT}}(n-6)n^{q-1}z^n}{Q_{\mathcal{E},K}(z,w)} \times \dfrac{1}{n-6}\left(\dfrac{(n-2)^{q-1}}{(n-2)^{q-1}+3n^{q-1}z^2}\right)} \\
&= \exp\left(\frac{U_{\mathcal{E}}(\omega)}{k_BT} - \frac{U_{\mathcal{E}}(\omega')}{k_BT}\right).
\end{aligned}$$

If $|\omega| = |\omega'| = n$ (*i.e.* ω' can be obtained from ω in one $p(0)$ Θ-BFACF move), and $\frac{\pi_{\omega'}(q,z,\mathcal{E})R_{\omega'\omega}}{\pi_{\omega}(q,z,\mathcal{E})R_{\omega\omega'}} < 1$, then it can be similarly shown that

$$\alpha_{\omega\omega'} = \exp\left(\frac{U_{\mathcal{E}}(\omega)}{k_B T} - \frac{U_{\mathcal{E}}(\omega')}{k_B T}\right).$$

Theorem 5.3.2. *If $R_{\omega\omega'}$ is the one-step transition probability of going from ω to ω' in one Θ-BFACF move, and $\alpha_{\omega\omega'}$ is the Metropolis sampling acceptance probability*

$$\alpha_{\omega\omega'} := \begin{cases} 1, & \text{if } R_{\omega\omega'} = 0, \\ \min\left(e^{\frac{U_{\mathcal{E}}(\omega)}{k_B T} - \frac{U_{\mathcal{E}}(\omega')}{k_B T}}, 1\right), & \text{if } R_{\omega\omega'} > 0, \end{cases} \tag{5.12}$$

then the Markov chain $\{X_t, t \in T\}$ on $\mathcal{P}^{\Theta}(K)$ defined by the one-step transition probabilities

$$P_{xy} = R_{xy}\alpha_{xy}, \ \forall\, x, y \in \mathcal{P}^{\Theta}(K), \tag{5.13}$$

is ergodic on $\mathcal{P}^{\Theta}(K)$ with the equilibrium distribution $\pi(q,z,\mathcal{E})$ as defined by Equation 5.8.

Proof. By the definition of P_{xy} and α_{xy} and the fact that the Θ-BFACF algorithm is ergodic on $\mathcal{P}^{\Theta}(K)$, Theorem 5.1.1 states that the corresponding Markov chain $\{X_t, t \in T\}$ defined by the one step transition probabilities P_{xy} is also reversible, aperiodic and irreducible on $\mathcal{P}^{\Theta}(K)$. By the definition of reversible (Definition 3.1.7), the Markov chain is positive recurrent. Because the Markov chain is positive recurrent, aperiodic and irreducible, by Definition 3.1.8 it is ergodic. Finally, also by Theorem 5.1.1,

$$\sum_{x \in \mathcal{P}^{\Theta}(K)} \pi_x(q,z,\mathcal{E})P_{xy} = \pi_y(q,z,\mathcal{E}), \ \forall\, y \in \mathcal{P}^{\Theta}(K). \tag{5.14}$$

Therefore, $\pi(q,z,\mathcal{E})$ is the equilibrium distribution of the Markov chain.

□

5.4 Definition of the I-Θ-BFACF Algorithm

Given a starting Θ-SAP $\omega^{[0]} \in P^{\Theta}(K)$, energy parameters defined by $\mathcal{E} = \{A, \zeta, v, T\}$, a fugacity $0 < z < z_c^{\Theta,\mathcal{E}}(K)$, and an integer $q > 0$, $t = 0$, and $X_0 = \omega^{[0]}$, the Markov chain $\{X_t, t \in T\}$ defined by this algorithm proceeds as follows:

Number the edges and vertices of X_t according to the rules outlined in the Θ-BFACF algorithm.

Choose one of the edges of X_t that is not in the Θ-structure uniformly at random, call this edge e^*.

Consider the embeddings W_1, W_2, W_3, and W_4 that result from moving e^* one unit in the 4 directions perpendicular to e^*, and then adding the necessary edges to join the newly shifted edge to the Θ-SAP, removing any double edges that might result.

For each $i = 1, 2, 3, 4$, define the probability of proposing the embedding W_i to be:

$$\Pr(W_i) := \left\{ \begin{array}{ll} \frac{(n+2)^{q-1}z^2}{n^{q-1}+3(n+2)^{q-1}z^2} =: p_n(+2), & \text{if } |W_i| - |X_t| = 2, \\ \frac{(n-2)^{q-1}}{(n-2)^{q-1}+3n^{q-1}z^2} =: p_n(-2), & \text{if } |W_i| - |X_t| = -2, \\ \frac{p_n(+2)+p_n(-2)}{2} =: p_n(0), & \text{if } |W_i| - |X_t| = 0. \end{array} \right\}. \tag{5.15}$$

Define $1 - \sum_{i=1}^{4} \Pr(W_i) = \Pr(X_t)$ to be the probability of doing nothing and seting $X_{t+1} = X_t$. Choose one of the 5 embeddings W_1, W_2, W_3, W_4, X_t according to their respective probabilities, and denote this chosen embedding to be W.

If W is not a Θ-SAP (*i.e.* it is not self-avoiding or the Θ-structure in X_t was altered), set $X_{t+1} = X_t$.

If W is a Θ-SAP, accept $X_{t+1} = W$ with probability $\alpha_{X_t W}$ and reject it (*i.e.* set $X_{t+1} = X_t$) with probability $1 - \alpha_{X_t W}$.

Increment t by 1 and repeat the above procedure.

5.5 How to Choose z-values

When simulating a Markov chain using the Θ-BFACF algorithm (with knot type K) in the good solvent case, one must choose a fugacity z such that $0 < z < z_c(K)$. In [46], Orlandini *et al.* estimate $\frac{1}{z_c(K)}$ for some simple knot types, namely,

$$\frac{1}{z_c(K)} = \left\{ \begin{array}{ll} 4.6852, & \text{if } K = \phi, \\ 4.6832, & \text{if } K = 3_1, \\ 4.6833, & \text{if } K = 4_1, \end{array} \right. \tag{5.16}$$

where these estimates are accurate up to the second decimal place. It is believed that $z_c(\phi) = z_c(3_1) = z_c(4_1)$ [45]; these estimates do not rule out that possibility. In order to obtain larger polygons in a chain, one must use values of z that are closer and closer to $z_c(K)$. Having a reasonable bound on this critical value, such as those provided in Equation 5.16 is helpful as it provides an upper bound for the z-values to consider.

When using the I-Θ-BFACF algorithm based on the energy parameters defined by $\mathcal{E} = \{\zeta, A, v, T\}$, one must now choose a fugacity z such that $0 < z < z_c^{\mathcal{E}}(K)$. This creates difficulties because the critical value $z_c^{\mathcal{E}}(K)$ varies depending on the energy parameters being used. As there is no literature that this author could find with results pertaining to estimates of $z_c^{\mathcal{E}}(K)$ or average equilibrium polygon lengths for different energy parameters (aside from the good solvent case and the $A = 0$ case), it was an "adventure" trying to find z-values that would yield polygons of a reasonable length at equilibrium for some choices of $\mathcal{E}$. By observing which simulations appeared to converge or diverge, I was able to get a rough idea of where the critical value was for each distribution. However, because some of the distributions considered here took a week or longer of computing time to converge (or diverge), it was very time consuming in some cases to find a z-value that would yield a large enough average equilibrium polygon length.

In the initial testing phase, q was set to be 1. For smaller values of ζ with $A/k_BT = 0.01$ and $v = -0.26$ fixed, there was so much difficulty trying to find a z-value that would converge to a distribution with a reasonable average equilibrium polygon length that q was increased to 2; it is possible that this problem with $q = 1$ for small values of ζ is similar to the issues that arose in the BFACF algorithm with $q = 1$ and $K = \phi$ (described in Section 4.2). Even with this increase in q, it was still difficult to find z-values that yielded a 'reasonable' average equilibrium polygon length (in this thesis, a reasonable average equilibrium length is around 200-400) for values of ζ smaller than 0.2. It was noticed that for these small values of ζ, the polygons in the chains with the highest z-values will often 'take off' to quite large lengths (2000 or greater); this creates two major problems. First of all, it takes a large amount of time steps for the lengths of polygons in the chain to come back down to average. Secondly, the I-Θ-BFACF algorithm takes up to $O(n)$ time (discussed next in Section 5.6), where n is the length of the polygon in the chain. So, not only does it take a larger amount of time steps for the chain to come back down to normal, but it also takes a longer amount of computational time for those time steps to be completed.

From the experience of searching for z-values that would yield average lengths around 200 for small values of ζ, it seems that the best procedure is to first find z-values that yield a smaller

average length then slowly working up to larger average lengths. If one immediately tries to find z-values that yield larger average lengths, weeks if not months of time can be lost from chains that are stuck due to polygons spending a long amount of time in excessively larger states.

5.6 An Updating Scheme To Increase Runtime Efficiency

Let $\{X_t, t \in T\}$ represent the Markov chain that results from an implementation of the I-Θ-BFACF algorithm based on the energy parameters $\mathcal{E} = \{\zeta, A, v, T\}$. When a new state X_* is proposed by the algorithm at each time step t, one must calculate the change in energy that results between this proposed state and the current state X_t in order to calculate the Metropolis sampling acceptance probability $\alpha_{X_t X_*} = \min\left(e^{\frac{U(X_t)-U(X_*)}{k_B T}}, 1\right)$. Supposing one already knows the energy $U(X_t)$, then one could naively compute $U(X_*)$ from scratch in order to determine $\alpha_{X_t X_*}$. This calculation takes $\mathrm{O}(n^2)$ time, where n is the length of X_*. Since the algorithm is only making a local change to X_t, we present an easier method for calculating $\alpha_{X_t X_*}$.

Recall from Section 2.2.5 that the potential energy of a SAP ω with length n can be expressed as

$$U_{\zeta,A,T,v}(\omega) = C(\omega) k_B T v + D_{A,\zeta}(\omega), \tag{5.17}$$

where $C(\omega)$ is the number of contacts in ω,

$$D_{A,\zeta}(\omega) = \sum_{i<j\leq n} I((\mathbf{v}_i, \mathbf{v}_j) \notin E(\omega)) \frac{A e^{-\zeta r_{ij}(\omega)}}{r_{ij}(\omega)}, \tag{5.18}$$

and $r_{ij}(\omega)$ is the euclidean distance between the i^{th} and j^{th} vertices of ω. Substituting Equation 5.17 into the Metropolis sampling acceptance probability yields

$$\alpha_{X_t X_*} = \min\left(e^{\gamma^{(1)}_{X_t X_*} + \gamma^{(2)}_{X_t X_*}}, 1\right), \tag{5.19}$$

where

$$\gamma^{(1)}_{X_t X_*} = \left(C(X_t) - C(X_*)\right) v, \tag{5.20}$$

and

$$\gamma^{(2)}_{X_t X_*} = \frac{D(X_t) - D(X_*)}{k_B T}. \tag{5.21}$$

Therefore, one only needs to compute $\gamma^{(1)}_{X_t X_*}$ and $\gamma^{(2)}_{X_t X_*}$ in order to calculate $\alpha_{X_t X_*}$.

5.6.1 Determining $\gamma^{(1)}_{X_t X_*}$

Because a Θ-BFACF move only proposes a local change to X_t, there will only be a small difference between the number of contacts in X_t and X_*. It should be noted that the following procedure takes $O(1)$ time.

Going from left to right in Figure 5.1 (*i.e.* applying a $p(+2)$ move), if a vertex marked with an asterisk is occupied, then an additional contact will occur as a result of the $p(+2)$ move. There is also an additional contact that will be created between the original two vertices that used to be joined by an edge. These are the only locations where contacts can be introduced by a $p(+2)$ move. It should be noted that a $p(+2)$ move will never remove an existing contact. Thus, if x represents the number of vertices marked by asterisks in Figure 5.1 that are occupied, then $\gamma^{(1)}_{X_t X_*} = -(x+1)v$. On the other hand, if we are going from right to left in Figure 5.1 (*i.e.* a $p(-2)$ move is being performed), then these $x+1$ contacts will be lost. Because a $p(-2)$ move can never introduce a new contact, then for a $p(-2)$ move $\gamma^{(1)}_{X_t X_*} = (x+1)v$.

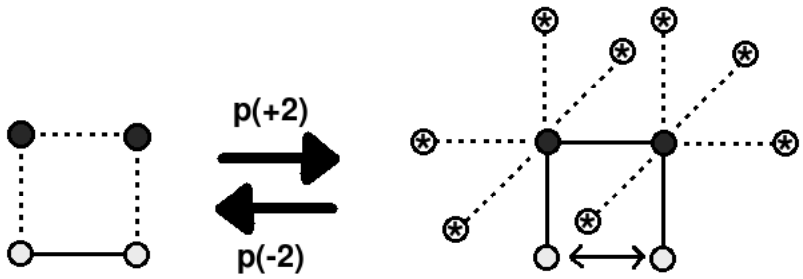

Figure 5.1: Each vertex with an asterisk that is occupied represents a contact that will be gained in a $p(+2)$ move or lost during a $p(-2)$ move. A $p(+2)$ move will also introduce an additional contact between the two vertices which were originally joined by an edge; similarly, this contact will be lost in a $p(-2)$ move.

The $p(0)$ move case is only slightly more complicated. In Figure 5.2, all the asterisk-marked vertices that are occupied represent a contact that will be lost as a result of the $p(0)$ move. Similarly, all the X-marked vertices that are occupied represent a contact that will be gained as a result of the $p(0)$ move. Thus, if x represents the number of asterisk-marked vertices that are occupied and y represents the number of X-marked vertices that are occupied, then $\gamma^{(1)}_{X_t X_*} = (x-y)v$.

5.6.2 Determining $\gamma^{(2)}_{X_t X_*}$

Recall from Equation 5.21 that

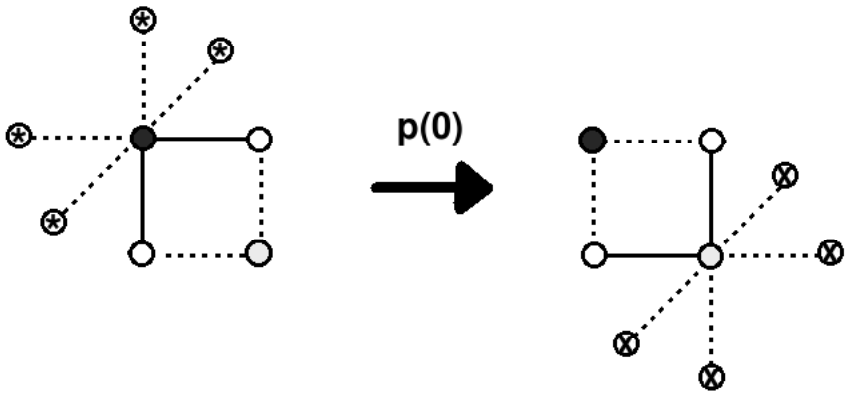

Figure 5.2: In the $p(0)$ move shown, each vertex with an asterisk that is occupied represents a contact that will be lost. Each vertex with an 'X' that is occupied represents a contact that will be gained due to the $p(0)$ move.

$$\gamma^{(2)}_{X_t X_*} = \frac{D(X_t) - D(X_*)}{k_B T} =: D'(X_t) - D'(X_*). \tag{5.22}$$

Because of the local nature of Θ-BFACF moves, many of the terms that are in $D'(X_t)$ will also be in $D'(X_*)$. The following procedure explains how to calculate $D'(X_t) - D'(X_*)$ in $O(n)$ time, where n is the length of X_t.

Suppose that the vertices in X_t are $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, and suppose that the state X_* is proposed from X_t by a $p(+2)$ move. Three new terms that are in $D'(X_*)$ are related to the interactions indicated by the arrows shown in Figure 5.3. For each of the two vertices that are added from the $p(+2)$ move (indicated by 'X's in Figure 5.3), there are interactions with the (n-2) vertices that are not part of the edge used in the $p(+2)$ move; these vertices are indicated by asterisks in Figure 5.3. Thus, only $2n-1$ interactions need to be calculated to determine $\gamma^{(2)}_{X_t X_*}$. If S represents the sum of all of these interactions, then $\gamma^{(2)}_{X_t X_*} = -S$. In the case of a $p(-2)$ move, all of these interactions are lost, and thus $\gamma^{(2)}_{X_t X_*} = S$.

Figure 5.4 shows an example of a $p(0)$ move. The interactions that occur between the vertex marked by a '-' and each vertex marked by an asterisk will be 'lost' as a result of the $p(0)$ move. Similarly, the interactions that occur between the vertex marked by a '+' and each vertex marked by an asterisk will be gained by the $p(0)$ move. If S^- represents the sum of the interactions between the vertex marked by a '-' and each vertex marked with an asterisk, and S^+ represents the sum of the interactions between the vertex marked by a '+' and each vertex marked with an asterisk, then

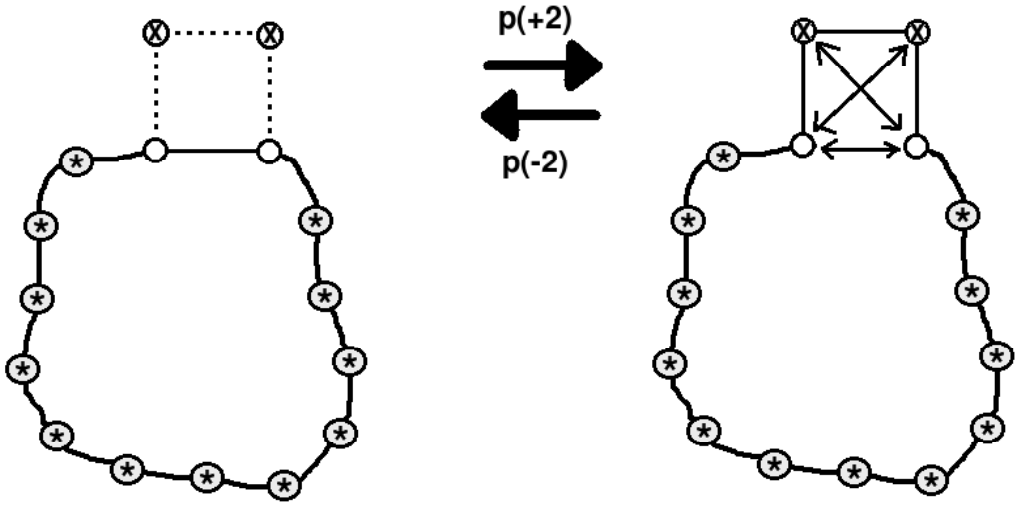

Figure 5.3: Three new interactions between vertices that result from a $p(+2)$ move are indicated by arrows in the right image. Also, each vertex marked with an 'X' has an interaction with each vertex marked with an asterisk. In the case of a $p(-2)$ move, all of these interactions mentioned are removed.

$\gamma^{(2)}_{X_t X_*} = S^- - S^+$. Thus, one only needs to calculate $2(n-3)$ interactions in this case in order to update $\gamma^{(2)}_{X_t X_*}$ (a linear time operation).

Using the constant time procedure in Section 5.6.1 to calculate $\gamma^{(1)}_{X_t X_*}$ and the linear time procedure in Section 5.6.2 to calculate $\gamma^{(1)}_{X_t X_*}$, one can calculate $\alpha_{X_t X_*}$ in $O(n)$ time. In order to ensure that this updating procedure yielded the correct change in energy for each time step, a short simulation of 10 million time steps was run where the energy was also calculated from scratch at each time step. At each time step t, the change in energy between states X_t and X_{t+1} was calculated using the updating procedure and compared to the actual difference of the energies of X_t and X_{t+1} calculated from scratch. Over the course of this simulation, the result from the updating procedure was always the same as the true difference in energy. Thus, the use of this updating procedure is reliable and represents considerable savings in CPU time compared to naively calculating the energy from scratch at each time step ($O(n^2)$ time).

5.7 Chapter Summary

Using Metropolis sampling based on the energy of a SAP as defined in Equation 2.26, it is possible to modify the pivot and Θ-BFACF algorithms to have relevant equilibrium distributions relating to specific solvent conditions. The pivot algorithm with Metropolis sampling based on SAP energy is

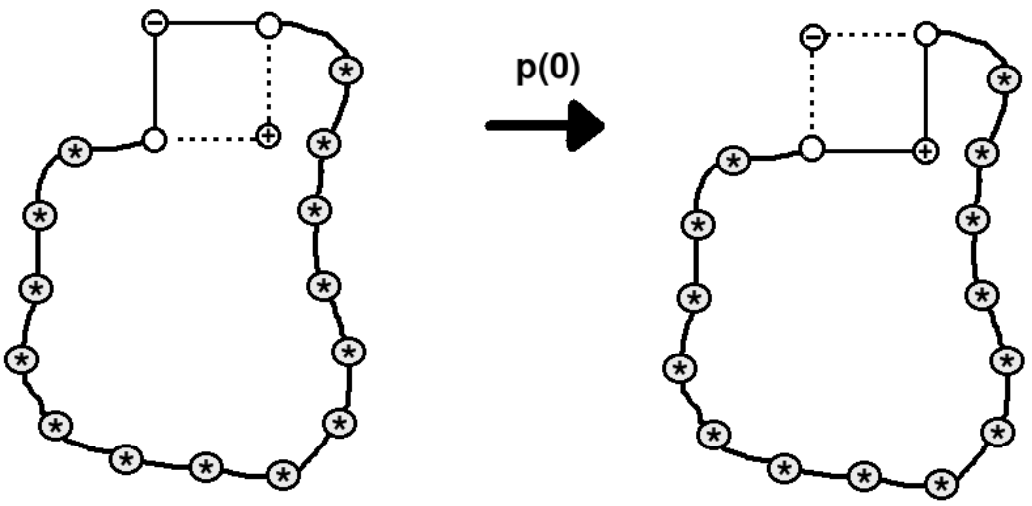

Figure 5.4: The interactions that occur between the vertex marked by a '-' and each vertex marked by an asterisk will no longer occur as a result of the $p(0)$ move; but there will be new interactions between the vertex marked by a '+' and each vertex marked by an asterisk.

referred to here as the *I-Pivot Algorithm*, and was first used by Tesi *et al.* in [64]. The Θ-BFACF algorithm with Metropolis sampling based on SAP is referred to here as the *I-Θ-BFACF Algorithm*, and is a new algorithm presented in this thesis. To generate equilibrium samples, the pivot or Θ-BFACF algorithm proceeds as it would in the case where a good solvent is assumed, with one step added at the end. Assuming the current state of the chain is x and the algorithm proposes y to be the next state in the chain, y is no longer automatically accepted as the next state in the chain; rather, it is accepted with the acceptance rate α_{xy}.

It was proved in this chapter that the new I-Θ-BFACF algorithm is ergodic, and that the radius of convergence of $Q^{\Theta}_{K,\mathcal{E}}(z,w)$ is positive for any proper choice of energy parameters (*i.e.* $A \geq 0$, $T \geq 0$, $v \leq 0$ and $\zeta \geq 0$. A new technique was also introduced which demonstrates how to increase the efficiency of the I-Θ-BFACF algorithm from $O(n^2)$ to $O(n)$ time, where n is the length of the polygon in the chain.

Now that all the necessary information has been provided for using Markov chains to sample from distributions that depend on the salt concentration of the solution, the only step left is to introduce the techniques necessary to analyze the data generated using these algorithms.

Chapter 6

Techniques for Analyzing CMC Data

In Chapter 3 it was reviewed how to determine when a Markov Chain (or a CMC) has reached equilibrium, as well as how to determine when two datapoints in the chain were essentially independent. The following chapter reviews some techniques that can be used to analyze essentially independent data coming from the equilibrium distribution of a Markov Chain or CMC in order to generate statistics for quantities of interest. As was the case in Chapter 3, a majority of the methods described in this chapter were described and used by Szafron in [61]. These methods are being reviewed again here because I wrote my own code in C or R to implement these methods; also, I felt that some of the results from these methods are not obvious and should be included for completeness. These methods will be used extensively in the results presented in Chapters 7 and 8.

6.1 Generating Confidence Intervals Using Data Coming From Essentially Independent Batches

Let $\{f(X_t), t \in \{0, 1, \ldots, t_0\}\}$ be a stochastic process consisting of data coming from its equilibrium distribution π, assuming it exists. Suppose this data is partitioned into n batch means $f(Y_{1,b}), \ldots, f(Y_{n,b})$, where each batch consists of b datapoints and $f(Y_{j,b}) = b^{-1} \sum_{i=1}^{b} f(X_{(j-1)b+i})$. Further, suppose these batches have passed the test for independence described in Section 3.2.6 at the $\alpha = 0.05$ level of significance. Because all the batch means are essentially independent, they can be treated as realizations of *i.i.d.* random variables. Therefore, a $(1-\alpha) \times 100\%$ confidence interval for $E_\pi(f)$ is:

$$\bar{f}(Y) \pm t_{n-1}(1-\alpha/2)\sqrt{\frac{s_Y}{n}}, \tag{6.1}$$

where $\bar{f}(Y)$ and s_Y are the respective average and standard deviation of the batch means $f(Y_{1,b}), \ldots, f(Y_{n,b})$, and $t_{n-1}(1-\alpha/2)$ is the $(1-\alpha/2)$ critical value of the t-distribution with $n-1$ degrees of freedom.

6.2 Ratio Estimation

Suppose that we have an ergodic Markov chain $\mathcal{X}$ that was started it in its equilibrium distribution π, and suppose that we are interested in estimating the observable f, where f is the ratio of two random variables g and h. The knot transition probabilities of Equation 2.9 are examples of quantities considered here for which ratio estimation is necessary. The following section describes a ratio estimation technique for such a ratio. This discussion is based on [61, Appendix A.3], and is included here for completeness.

Suppose that $\{(X_i, Y_i), i = 1, \ldots, n\}$ is a sequence of independent, indentically distributed random two-dimensional vectors with $\mu_Y := E[Y_i]$, $\mu_X := E[X_i] \neq 0$, $\sigma_Y^2 := E[(Y_i - \mu_Y)^2] < \infty$, $\sigma_X^2 := E[(X_i - \mu_X)^2] < \infty$, and $\sigma_{X,Y}^2 := E[(X_i - \mu_X)(Y_i - \mu_Y)] < \infty$ for $i = 1, \ldots, n$. Define $\theta := \frac{\mu_Y}{\mu_X}$ and

$$\bar{\theta}_n := \begin{cases} \frac{\bar{Y}_n}{\bar{X}_n}, & \text{if } \bar{X}_n \neq 0 \\ 0, & \text{otherwise,} \end{cases} \tag{6.2}$$

where

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \text{ and } \bar{Y}_n := \frac{1}{n}\sum_{i=1}^{n} Y_i. \tag{6.3}$$

The following theorems are proved by Fishman in [16]:

Theorem 6.2.1 ([16])**.**

$$\lim_{n\to\infty} nE\left[\bar{\theta}_n - \theta\right] = \theta\left[\frac{\sigma_X^2}{\mu_X^2} - \frac{\sigma_{X,Y}^2}{\mu_X\mu_Y}\right]. \tag{6.4}$$

Theorem 6.2.2 ([16])**.**

$$\lim_{n\to\infty} nE\left[(\bar{\theta}_n - \theta)^2\right] = \theta^2\left[\frac{\sigma_X^2}{\mu_X^2} - 2\frac{\sigma_{X,Y}^2}{\mu_X\mu_Y} + \frac{\sigma_Y^2}{\mu_Y^2}\right]. \tag{6.5}$$

Theorem 6.2.1 shows that $\hat{\theta}$ is a biased estimator of θ. To reduce this bias, Fishman [16] recommends using the estimator

$$\tilde{\theta}_n := \bar{\theta}_n\left[1 + \frac{1}{n}\left(\frac{\hat{\sigma}_{X,Y}^2}{\bar{X}_n\bar{Y}_n} - \frac{\hat{\sigma}_X^2}{\bar{X}_n^2}\right)\right], \tag{6.6}$$

where $\hat{\sigma}_X^2$ and $\hat{\sigma}_Y^2$ are the sample variances for X and Y and $\hat{\sigma}_{X,Y}^2$ is the sample covariance of X and Y. This recommendation is a consequence of the following theorem [68]:

Theorem 6.2.3 ([68]). 1. $\lim_{n\to\infty} nE\left[\tilde{\theta}_n - \theta\right] = 0$, *and*
2. $\lim_{n\to\infty} nE\left[(\tilde{\theta}_n - \theta)^2\right] = \theta^2\left[\frac{\sigma_X^2}{\mu_X^2} - 2\frac{\sigma_{X,Y}^2}{\mu_X\mu_Y} + \frac{\sigma_Y^2}{\mu_Y^2}\right]$.

Thus, $\tilde{\theta}_n$ can be considered an essentially unbiased estimator for large enough n, and there is no additional cost in variance compared to $\bar{\theta}_n$. Hence, define the estimator for the variance of $\tilde{\theta}$ to be

$$\hat{\text{var}}(\tilde{\theta}_n) := \frac{\tilde{\theta}_n}{n}\left[\frac{\hat{\sigma}_X^2}{\bar{X}_n^2} + \frac{\hat{\sigma}_Y^2}{\bar{Y}_n^2} - \frac{2\hat{\sigma}_{X,Y}^2}{\bar{X}_n\bar{Y}_n}\right]. \tag{6.7}$$

In order to determine a 95% confidence interval for θ, define

$$V_i := Y_i - \theta X_i, \;\; i = 1, \ldots, n, \tag{6.8}$$

and

$$\bar{V}_n = \bar{Y}_n - \theta\bar{X}_n. \tag{6.9}$$

Because $E[\bar{V}_n] = 0$,

$$\text{var}(\bar{V}_n) := E[(\bar{V}_n - E[\bar{V}_n])^2] = \left(\theta^2\sigma_X^2 - 2\theta\sigma_{X,Y}^2 + \sigma_Y^2\right)/n, \tag{6.10}$$

$$\text{var}(V_i) = n\text{var}(\bar{V}_n), \tag{6.11}$$

and

$$\hat{\text{var}}(V) := \theta^2\hat{\sigma}_X^2 - 2\theta\hat{\sigma}_{X,Y}^2 + \hat{\sigma}_Y^2. \tag{6.12}$$

Since $V_1, \ldots, V_n$ are *i.i.d.* random variables, the Central Limit Theorem yields the result that $\frac{\bar{V}_n}{\sqrt{\text{var}(\bar{V}_n)}}$ is asymptotically distributed as standard normal. Fishman proves that $\frac{\bar{V}_n}{\sqrt{\hat{\text{var}}(V)/n}}$ also has an asymptotic standard normal distribution, and for large values of n,

$$\Pr\left[\frac{|\bar{V}_n|}{\sqrt{\hat{\text{var}}(V)/n}} \le c(\alpha)\right] \approx 1 - \alpha, \tag{6.13}$$

where $c(\alpha)$ is the value for which

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{c(\alpha)} e^{-z^2/2}dz = 1 - \alpha/2, \text{ for } 0 < \alpha < 1. \tag{6.14}$$

Because

$$\Pr\left[\frac{|\bar{V}_n|}{\sqrt{\hat{\text{var}}(V)/n}} \le c(\alpha)\right] = \Pr\left[|\bar{Y}_n - \theta\bar{X}_n| \le c(\alpha)\sqrt{\frac{\theta^2\hat{\sigma}_X^2 - 2\theta\hat{\sigma}_{X,Y}^2 + \hat{\sigma}_Y^2}{n}}\right], \tag{6.15}$$

$$(\bar{Y}_n - \theta\bar{X}_n)^2 \leq \frac{c^2(\alpha)}{n}(\theta^2\hat{\sigma}_X^2 - 2\theta\hat{\sigma}_{X,Y}^2 + \hat{\sigma}_Y^2). \tag{6.16}$$

Expanding the left hand side of Equation 6.16 and simplifying yields the following quadratic inequality in θ:

$$\left[\bar{X}_n^2 - \frac{c^2(\alpha)}{n}\hat{\sigma}_X^2\right]\theta^2 - 2\theta\left[\bar{X}_n\bar{Y}_n - \frac{c^2(\alpha)}{n}\hat{\sigma}_{X,Y}^2\right] + \left[\bar{Y}_n^2 - \frac{c^2(\alpha)}{n}\hat{\sigma}_Y^2\right] \leq 0. \tag{6.17}$$

Solving the previous quadratic inequality, provided real solutions to it exist, yields the interval $r_1 \leq \theta \leq r_2$ for θ, where

$$r_1 := \frac{\bar{X}_n\bar{Y}_n - \frac{c^2(\alpha)}{n}\hat{\sigma}_{X,Y}^2 - \sqrt{\left[\bar{X}_n\bar{Y}_n - \frac{c^2(\alpha)}{n}\hat{\sigma}_{X,Y}^2\right]^2 - \left[\bar{X}_n^2 - \frac{c^2(\alpha)}{n}\hat{\sigma}_X^2\right]\left[\bar{Y}_n^2 - \frac{c^2(\alpha)}{n}\hat{\sigma}_Y^2\right]}}{\bar{X}_n^2 - \frac{c^2(\alpha)}{n}\hat{\sigma}_X^2}, \tag{6.18}$$

and

$$r_2 := \frac{\bar{X}_n\bar{Y}_n + \frac{c^2(\alpha)}{n}\hat{\sigma}_{X,Y}^2 - \sqrt{\left[\bar{X}_n\bar{Y}_n - \frac{c^2(\alpha)}{n}\hat{\sigma}_{X,Y}^2\right]^2 - \left[\bar{X}_n^2 - \frac{c^2(\alpha)}{n}\hat{\sigma}_X^2\right]\left[\bar{Y}_n^2 - \frac{c^2(\alpha)}{n}\hat{\sigma}_Y^2\right]}}{\bar{X}_n^2 - \frac{c^2(\alpha)}{n}\hat{\sigma}_X^2}. \tag{6.19}$$

Whenever $r_1, r_2 \in \mathbb{R}$, the interval $r_1 \leq \theta \leq r_2$ is a $(1-\alpha)\times 100\%$ confidence interval for θ.

6.2.1 Ratio Estimation using CMC data

Suppose $\{W_t := (W_t(1), \ldots, W_t(M)), t \in \{1, 2, \ldots t_0\}\}$ is a composite Markov chain on the state space $\mathcal{S}^M$ coming from its equilibrium distribution. Further, suppose that X and Y are real valued observable functions defined on $\mathcal{S}$ such that $\theta = \frac{\mu_X}{\mu_Y}$, where $\mu_Y := E[Y]$, $\mu_X := E[X] \neq 0$, and the variances and covariance of X and Y are finite. Define

$$X(W_t) := (X(W_t(1)), \ldots, X(W_t(M))) \tag{6.20}$$

and

$$Y(W_t) := (Y(W_t(1)), \ldots, Y(W_t(M))) \tag{6.21}$$

to be the realizations of X and Y for each chain in the CMC at time step t. Suppose that τ_{int} is known, and that there are $\lfloor t_0/2\tau_{\text{int}} \rfloor := n_B$ essentially independent blocks of data. Suppose now

that we take a subsample in block k by selecting every r^{th} point, thus getting a subsample of length $\lfloor 2\tau_{\text{int}}/r \rfloor := n_s$. Say the subsample for block k is denoted by ω_k, and consists of the n_s datapoints $\omega_{k,1}, \ldots, \omega_{k,n_s}$. For a fixed block k, let $X_{k,i}$ be the sum of $X(\omega_{k,j}(i))$ over all $j = 1, \ldots, n_s$. Similarly, let $Y_{k,i}$ be the sum of $Y(\omega_{k,j}(i))$ over all $j = 1, \ldots, n_s$.

The definition of $X_{k,i}$ and $Y_{k,i}$ can be stated more formally; define

$$X_{k,i} := \sum_{t=1}^{t_0} I\left(t \in \left[2(k-1)\tau_{\text{int}} + 1, 2k\tau_{\text{int}}\right]\right) I\left(t \bmod r = 0\right) X(W_t(i)) \tag{6.22}$$

and

$$Y_{k,i} := \sum_{t=1}^{t_0} I\left(t \in \left[2(k-1)\tau_{\text{int}} + 1, 2k\tau_{\text{int}}\right]\right) I\left(t \bmod r = 0\right) Y(W_t(i)), \tag{6.23}$$

where I is the indicator function; in this context

$$I(a \in b) = \begin{cases} 1, & \text{if } a \in b, \\ 0, & \text{if } a \notin b, \end{cases} \tag{6.24}$$

and

$$I(a = b) = \begin{cases} 1, & \text{if } a = b, \\ 0, & \text{if } a \neq b. \end{cases} \tag{6.25}$$

Define the functions

$$X_{k,\cdot} := \frac{1}{M} \sum_{i=1}^{M} X_{k,i} \tag{6.26}$$

and

$$Y_{k,\cdot} := \frac{1}{M} \sum_{i=1}^{M} Y_{k,i}. \tag{6.27}$$

A point estimate for θ that is only based on chain i uses the sequence

$$\left((X_{k,i}, Y_{k,i}\right), k = 1, \ldots, n_B) \tag{6.28}$$

for ratio estimation in Equation 6.6, whereas if the point estimate is based on data in all the chains, the sequence

$$\left((X_{k,\cdot}, Y_{k,\cdot}\right), k = 1, \ldots, n_B) \tag{6.29}$$

is used in Equation 6.6.

6.3 Reliable Data - the choice of $\mathbf{N_{max}}(*)$

The following section is based on the discussion presented in [61, Section 4.6].

Because a simulation is finite, the observed proportions of large polygons may not accurately reflect the corresponding proportions determined using the true distribution. Suppose a simulation consists of n_0 replications, where for replication $r \in \{1, \ldots, n_0\}$, $\hat{g}_{2n}^{(r)}(*)$ is the estimated value of some observable quantity of interest on polygons whose lengths are $2n$; suppose the estimated standard error of this estimate is $\hat{SE}(\hat{g}_{2n}^{(r)}(*))$. The estimated relative standard error of $\hat{g}_{2n}^{(r)}(*)$ is defined to be:

$$\hat{\delta}_{2n}^{(r)}(*) := \begin{cases} \dfrac{\hat{SE}(\hat{g}_{2n}^{(r)}(*))}{\hat{g}_{2n}^{(r)}(*)}, & \text{if } \hat{SE}(\hat{g}_{2n}^{(r)}(*)) \neq 0 \\ \infty, & \text{otherwise.} \end{cases} \tag{6.30}$$

Now define

$$\hat{\delta}^{(r)}(*) := \min_n \hat{\delta}_{2n}^{(r)}(*), \tag{6.31}$$

and define $\hat{\eta}^{(r)}(*)$ to be the first value of $2n$ for which $\hat{\delta}_{2n}^{(r)}(*) = \hat{\delta}^{(r)}(*)$. Note that $\hat{\delta}^{(r)}(*)$ is the smallest relative error of $\hat{g}_{2n}^{(r)}(*)$ that can be achieved without generating more data. For a fixed amount of data, the most accurate data will be for values of n such that $\hat{\delta}_{2n}^{(r)}(*)$ is within some tolerance ϵ_* of $\hat{\delta}^{(r)}(*)$. How should ϵ_* be determined?

If $\epsilon_* > 1.0$, then the estimated error of the point estimate $\hat{g}_{2n}^{(r)}(*)$ would be greater than $\hat{g}_{2n}^{(r)}(*)$ itself. Also, any error in $\hat{g}_{2n}^{(r)}(*)$ would be introduced into subsequent calculations involving $\hat{g}_{2n}^{(r)}(*)$. Hence, having $\epsilon_* < 1.0$ is preferred. Define

$$\epsilon_* := \min_r(\hat{\delta}^{(r)}(*) + c), \tag{6.32}$$

where $c \times 100\%$ represents the maximum tolerated deviated from $\hat{\delta}^{(r)}(*)$ and c is chosen so that $0 < c < 1$ and $\epsilon_* < 1$. The choice of c is arbitrary; however, c should be chosen in such a manner that using the point estimates $\hat{g}_{2n}^{(r)}(*)$ whose estimated relative error is less than ϵ_*, minimizes the error introduced into subsequent calculations involving $\hat{g}_{2n}^{(r)}(*)$.

Assuming that c has been chosen, define $\hat{N}_{\max}(*)$ to be the first value of $2n > \hat{\eta}^{(r)}(*)$ for which $\hat{\delta}_{2n}^{(r)}(*)$ first achiveves a value greater than or equal to ϵ_* (over all n_0 replications). The set of

polygons whose lengths are greater than $\hat{N}_{\max}(*)$ will be referred to as *unreliable data*; the set of polygon lengths less than or equal to $\hat{N}_{\max}(*)$ will be referred to as *reliable data*. In determining $\hat{N}_{\max}(*)$, it was decided that a tolerance level of $c = 0.05$ would be used; this represents a 5% maximum tolerated deviation from $\hat{\delta}^{(r)}(*)$. This particular value of c was chosen in analysis done in [61]; upon examining the simulation data presented in Chapter 8, it was determined that a 5% deviation was not unreasonable.

6.4 Fixed-n analysis

This type of analysis can be used whenever one wants to fit a model to a series of estimates corresponding to varying polygon lengths. This type of analysis can be used to estimate limiting knot transition probabilities, the limiting probability of a successful strand passage, and the growth rate of the mean square radius of gyration.

Let X and Y be random variables defined on $\mathcal{S}$. Suppose that we are trying to estimate parameters $a_1, \ldots, a_k$ corresponding to some relationship between X and Y. Let

$$S := ((x_i, y_i), i = 1, \ldots, N) \tag{6.33}$$

be a sequence of observations of X and Y. In this work, each x_i will most likely correspond to some Θ-SAP length n_i, and y_i will be some function (e.g. knotting probabilities) estimated for that value of n_i. In order to estimate the parameters $a_1, \ldots, a_k$, an independent subsample from S is required [61]. This subsample is determined by finding the smallest k such that the points $H = \{(x_1, y_1), (x_{1+k}, y_{1+k}), \ldots, (x_{1+mk}, y_{1+mk})\}$ are independent, where $m = \lfloor \frac{N-1}{K} \rfloor$. Using only the $m + 1$ essentially independent datapoints in H, weighted least squares regression can be used to fit the data and estimate $a_1, \ldots, a_k$.

6.5 Grouped-n Analysis for I-Θ-BFACF Algorithm Data

A limitation of the fixed-n analysis technique is that most of the data has to be discarded. This seems wasteful; is there any way to be able to use more data than what comes from the set H of essentially independent datapoints? The answer to this question is "maybe". Before this question is fully addressed, we need to define some new terminology for strand passage statistics based on I-Θ-BFACF Algorithm simulations.

Given a CMC implementation of the I-Θ-BFACF algorithm with M chains where the energy parameters are $\mathcal{E} := \{A, v, \zeta, T\}$, q is a positive integer, and the fugacities for the M chains are $z_1, \ldots, z_M$, at equilibrium the probability of observing a Θ-SAP with length $2n$ in chain i is:

$$\Pr_{2n}^{\Theta,\mathcal{E},i}(\phi) := \frac{\sum_{\omega \in \mathcal{P}_{2n}^{\Theta}(\phi)} e^{\frac{-U_{\mathcal{E}}(\omega)}{k_B T}} (2n-6)(2n)^{q-1} z_i^{2n}}{\sum_{m=1}^{\infty} \sum_{\omega' \in \mathcal{P}_{2m}^{\Theta}(\phi)} e^{\frac{-U_{\mathcal{E}}(\omega')}{k_B T}} (2m-6)(2m)^{q-1} z_i^{2m}}. \tag{6.34}$$

If we let

$$Z_{2n}^{\Theta,\mathcal{E}}(\phi) := \sum_{\omega \in \mathcal{P}_{2n}^{\Theta}(\phi)} e^{\frac{-U_{\mathcal{E}}(\omega)}{k_B T}}, \tag{6.35}$$

$$w(n) := (n-6)n^{q-1}, \tag{6.36}$$

and

$$Q_{\phi,\mathcal{E}}^{\Theta}(z_i, w) = \sum_{m=1}^{\infty} \sum_{\omega' \in \mathcal{P}_{2m}^{\Theta}(\phi)} e^{\frac{-U_{\mathcal{E}}(\omega')}{k_B T}} (2m-6)(2m)^{q-1} z_i^{2m}, \tag{6.37}$$

Equation 6.34 simplifies to:

$$\Pr_{2n}^{\Theta,\mathcal{E},i}(\phi) = \frac{Z_{2n}^{\Theta,\mathcal{E}}(\phi) w(2n) z_i^{2n}}{Q_{\phi,\mathcal{E}}^{\Theta}(z_i, w)}. \tag{6.38}$$

Similarly, at equilibrium the probability of observing a length $2n$ Θ-SAP in chain i for which strand passage is successful is

$$\Pr_{2n}^{\Theta,\mathcal{E},i}(s|\phi) = \frac{Z_{2n}^{\Theta,\mathcal{E}}(s|\phi) w(2n) z_i^{2n}}{Q_{\phi,\mathcal{E}}^{\Theta}(z_i, w)}, \tag{6.39}$$

where

$$Z_{2n}^{\Theta,\mathcal{E}}(s|\phi) := \sum_{\omega \in \mathcal{P}_{2n}^{\Theta}(s|\phi)} e^{\frac{-U_{\mathcal{E}}(\omega)}{k_B T}}, \tag{6.40}$$

and the probability of observing a length $2n$ Θ-SAP in chain i for which strand passage is successful and the resulting knot type after strand passage is $K \in \mathcal{K}(\phi)$ is

$$\Pr_{2n}^{\Theta,\mathcal{E},i}(\phi \to K) = \frac{Z_{2n}^{\Theta,\mathcal{E}}(K|\phi, s) w(2n) z_i^{2n}}{Q_{\phi,\mathcal{E}}^{\Theta}(z_i, w)}, \tag{6.41}$$

where

$$Z_{2n}^{\Theta,\mathcal{E}}(K|\phi,s) := \sum_{\omega\in\mathcal{P}_{2n}^{\Theta}(K|\phi,s)} e^{\frac{-U_{\mathcal{E}}(\omega)}{k_BT}}. \tag{6.42}$$

Given that we have an unknotted Θ-SAP ω in chain i with length $2n$, at equilibrium the probability that we can perform a successful strand passage on ω is denoted by

$$\rho_{2n}^{\Theta,\mathcal{E}}(s|\phi) := \frac{\Pr_{2n}^{\Theta,\mathcal{E},i}(s|\phi)}{\Pr_{2n}^{\Theta,\mathcal{E},i}(\phi)} = \frac{Z_{2n}^{\Theta,\mathcal{E}}(s|\phi)}{Z_{2n}^{\Theta,\mathcal{E}}(\phi)}, \tag{6.43}$$

and the probability of a length $2n$ unknotted Θ-SAP going to knot type $K \in \mathcal{K}(\phi)$ given a successful strand passage is denoted by

$$\rho_{2n}^{\Theta,\mathcal{E}}(\phi \to K) := \frac{\Pr_{2n}^{\Theta,\mathcal{E},i}(\phi \to K)}{\Pr_{2n}^{\Theta,\mathcal{E},i}(s|\phi)} = \frac{Z_{2n}^{\Theta,\mathcal{E}}(K|s,\phi)}{Z_{2n}^{\Theta,\mathcal{E}}(s|\phi)}. \tag{6.44}$$

It is standard to assume that $Z_{2n}^{\Theta,\mathcal{E}}(\phi)$, $Z_{2n}^{\Theta,\mathcal{E}}(s|\phi)$, and $Z_{2n}^{\Theta,\mathcal{E}}(K|s,\phi)$ will all scale with n in the form:

$$\mathcal{G}_n(a,b,c,g,h) := an^b e^{cn}\left(1 + \frac{g}{n^h} + O(n^{-1})\right). \tag{6.45}$$

More specifically, define the scaling forms for $Z_{2n}^{\Theta,\mathcal{E}}(\phi)$, $Z_{2n}^{\Theta,\mathcal{E}}(s|\phi)$, and $Z_{2n}^{\Theta,\mathcal{E}}(K|s,\phi)$ to be

$$\mathcal{G}_n(A_{\phi}^{\Theta,\mathcal{E}}, \alpha_{\phi}^{\Theta,\mathcal{E}}, \kappa_{\phi}^{\Theta,\mathcal{E}}, B_{\phi}^{\Theta,\mathcal{E}}, \Delta_{\phi}^{\Theta,\mathcal{E}}), \tag{6.46}$$

$$\mathcal{G}_n(A_{s|\phi}^{\Theta,\mathcal{E}}, \alpha_{s|\phi}^{\Theta,\mathcal{E}}, \kappa_{s|\phi}^{\Theta,\mathcal{E}}, B_{s|\phi}^{\Theta,\mathcal{E}}, \Delta_{s|\phi}^{\Theta,\mathcal{E}}), \tag{6.47}$$

and

$$\mathcal{G}_n(A_{K|\phi,s}^{\Theta,\mathcal{E}}, \alpha_{K|\phi,s}^{\Theta,\mathcal{E}}, \kappa_{K|\phi,s}^{\Theta,\mathcal{E}}, B_{K|\phi,s}^{\Theta,\mathcal{E}}, \Delta_{K|\phi,s}^{\Theta,\mathcal{E}}), \tag{6.48}$$

respectively.

There is now sufficient notation to discuss the *grouped-n method* for estimating strand passage related probabilities. The following derivation is similar to the derivation presented in [61, p. 262] for a CMC implementation of the Θ-BFACF algorithm in the good solvent case. The only difference between the derivation here and the derivation presented in [61] is that anywhere there is a $Z_n^{\Theta,\mathcal{E}}(*)$ here it replaces a $p_n^{\Theta}(*)$ term in [61]. This argument was included as the result is not necessarily obvious.

Given even values n_1 and n_2 such that $n_1 < n_2$, define the *grouped-*$[n_1, n_2]$ *probability* of successful strand passage to be:

$$\rho^{\Theta,\mathcal{E}}_{[n_1,n_2]}(s|\phi) := \frac{\displaystyle\sum_{n=n_1}^{n_2}\left[Z_n^{\Theta,\mathcal{E}}(s|\phi)\sum_{i=1}^{M}\frac{w(n)z_i^n}{Q^{\Theta}_{\phi,\mathcal{E}}(z_i,w)}\right]}{\displaystyle\sum_{n=n_1}^{n_2}\left[Z_n^{\Theta,\mathcal{E}}(\phi)\sum_{i=1}^{M}\frac{w(n)z_i^n}{Q^{\Theta}_{\phi,\mathcal{E}}(z_i,w)}\right]}, \tag{6.49}$$

and the grouped-$[n_1, n_2]$ probability of obtaining a knot type $K \in \mathcal{K}(\phi)$ given a successful strand passage to be

$$\rho^{\Theta,\mathcal{E}}_{[n_1,n_2]}(\phi \to K) := \frac{\displaystyle\sum_{n=n_1}^{n_2}\left[Z_n^{\Theta,\mathcal{E}}(K|\phi,s)\sum_{i=1}^{M}\frac{w(n)z_i^n}{Q^{\Theta}_{\phi,\mathcal{E}}(z_i,w)}\right]}{\displaystyle\sum_{n=n_1}^{n_2}\left[Z_n^{\Theta,\mathcal{E}}(s|\phi)\sum_{i=1}^{M}\frac{w(n)z_i^n}{Q^{\Theta}_{\phi,\mathcal{E}}(z_i,w)}\right]}, \tag{6.50}$$

where both sums are taken through even values of n. Then $\rho^{\Theta,\mathcal{E}}_{[n_1,n_2]}(s|\phi)$ is the probability of observing a successful strand passage Θ-SAP given that it has a length somewhere in $[n_1, n_2]$, and $\rho^{\Theta,\mathcal{E}}_{[n_1,n_2]}(\phi \to K)$ is the probability of observing a knot type K given a successful strand passage of a Θ-SAP whose length is in $[n_1, n_2]$. The goal is to show that these grouped n probabilities have the same asymptotic behaviour as fixed n probabilities as n_1 tends to ∞.

$\rho^{\Theta,\mathcal{E}}_{[n_1,n_2]}(s|\phi)$ can be algebraically manipulated to be

$$\rho^{\Theta,\mathcal{E}}_{[n_1,n_2]}(s|\phi) = \frac{Z_{n_1}^{\Theta,\mathcal{E}}(s|\phi)}{Z_{n_1}^{\Theta,\mathcal{E}}(\phi)}\frac{\displaystyle\sum_{n=n_1}^{n_2}\left[\frac{Z_n^{\Theta,\mathcal{E}}(s|\phi)}{Z_{n_1}^{\Theta,\mathcal{E}}(s|\phi)}\sum_{i=1}^{M}\frac{w(n)z_i^n}{Q^{\Theta}_{\phi,\mathcal{E}}(z_i,w)}\right]}{\displaystyle\sum_{n=n_1}^{n_2}\left[\frac{Z_n^{\Theta,\mathcal{E}}(\phi)}{Z_{n_1}^{\Theta,\mathcal{E}}(\phi)}\sum_{i=1}^{M}\frac{w(n)z_i^n}{Q^{\Theta}_{\phi,\mathcal{E}}(z_i,w)}\right]}, \tag{6.51}$$

and if we substitute the believed scaling forms for $Z_n^{\Theta,\mathcal{E}}(\phi)$ and $Z_n^{\Theta,\mathcal{E}}(s|\phi)$ given by Equations 6.46 and 6.47, the above equation yields the following scaling form as $n_1 \to \infty$, up to first order:

$$\rho^{\Theta,\mathcal{E}}_{[n_1,n_2]}(s|\phi) \sim \frac{Z^{\Theta,\mathcal{E}}_{n_1}(s|\phi)}{Z^{\Theta,\mathcal{E}}_{n_1}(\phi)} \frac{\displaystyle\sum_{n=n_1}^{n_2} \left[\frac{n^{\alpha^{\Theta,\mathcal{E}}_{s|\phi}} e^{\kappa^{\Theta,\mathcal{E}}_{s|\phi} n} \left(1 + \frac{B^{\Theta,\mathcal{E}}_{s|\phi}}{n^{\Delta^{\Theta,\mathcal{E}}_{s|\phi}}}\right)}{n_1^{\alpha^{\Theta,\mathcal{E}}_{s|\phi}} e^{\kappa^{\Theta,\mathcal{E}}_{s|\phi} n_1} \left(1 + \frac{B^{\Theta,\mathcal{E}}_{s|\phi}}{n_1^{\Delta^{\Theta,\mathcal{E}}_{s|\phi}}}\right)} \sum_{i=1}^{M} \frac{w(n) z_i^n}{Q^{\Theta}_{\phi,\mathcal{E}}(z_i, w)} \right]}{\displaystyle\sum_{n=n_1}^{n_2} \left[\frac{n^{\alpha^{\Theta,\mathcal{E}}_{\phi}} e^{\kappa^{\Theta,\mathcal{E}}_{\phi} n} \left(1 + \frac{B^{\Theta,\mathcal{E}}_{\phi}}{n^{\Delta^{\Theta,\mathcal{E}}_{\phi}}}\right)}{n_1^{\alpha^{\Theta,\mathcal{E}}_{\phi}} e^{\kappa^{\Theta,\mathcal{E}}_{\phi} n_1} \left(1 + \frac{B^{\Theta,\mathcal{E}}_{\phi}}{n_1^{\Delta^{\Theta,\mathcal{E}}_{\phi}}}\right)} \sum_{i=1}^{M} \frac{w(n) z_i^n}{Q^{\Theta}_{\mathcal{E},\phi}(z_i, w)} \right]}. \tag{6.52}$$

If we assume that $\alpha^{\Theta,\mathcal{E}}_{\phi} = \alpha^{\Theta,\mathcal{E}}_{s|\phi}$ and $\kappa^{\Theta,\mathcal{E}}_{\phi} = \kappa^{\Theta,\mathcal{E}}_{s|\phi}$, then the scaling form in Equation 6.52 can be 'simplified' to be:

$$\rho^{\Theta,\mathcal{E}}_{[n_1,n_2]}(s|\phi) \sim \frac{Z^{\Theta,\mathcal{E}}_{n_1}(s|\phi)}{Z^{\Theta,\mathcal{E}}_{n_1}(\phi)} \frac{\left(1 + \frac{B^{\Theta,\mathcal{E}}_{\phi}}{n_1^{\Delta^{\Theta,\mathcal{E}}_{\phi}}}\right)}{\left(1 + \frac{B^{\Theta,\mathcal{E}}_{s|\phi}}{n_1^{\Delta^{\Theta,\mathcal{E}}_{s|\phi}}}\right)} \frac{\displaystyle\sum_{n=n_1}^{n_2} \left[n^{\alpha^{\Theta,\mathcal{E}}_{s|\phi}} e^{\kappa^{\Theta,\mathcal{E}}_{s|\phi} n} \left(1 + \frac{B^{\Theta,\mathcal{E}}_{s|\phi}}{n^{\Delta^{\Theta,\mathcal{E}}_{s|\phi}}}\right) \sum_{i=1}^{M} \frac{w(n) z_i^n}{Q^{\Theta}_{\phi,\mathcal{E}}(z_i, w)} \right]}{\displaystyle\sum_{n=n_1}^{n_2} \left[n^{\alpha^{\Theta,\mathcal{E}}_{\phi}} e^{\kappa^{\Theta,\mathcal{E}}_{\phi} n} \left(1 + \frac{B^{\Theta,\mathcal{E}}_{\phi}}{n^{\Delta^{\Theta,\mathcal{E}}_{\phi}}}\right) \sum_{i=1}^{M} \frac{w(n) z_i^n}{Q^{\Theta}_{\phi,\mathcal{E}}(z_i, w)} \right]}. \tag{6.53}$$

If we define the variables

$$A_1 := \frac{Z^{\Theta,\mathcal{E}}_{n_1}(s|\phi)}{Z^{\Theta,\mathcal{E}}_{n_1}(\phi)}, \tag{6.54}$$

$$A_2 := \frac{\left(1 + \frac{B^{\Theta,\mathcal{E}}_{\phi}}{n_1^{\Delta^{\Theta,\mathcal{E}}_{\phi}}}\right)}{\left(1 + \frac{B^{\Theta,\mathcal{E}}_{s|\phi}}{n_1^{\Delta^{\Theta,\mathcal{E}}_{s|\phi}}}\right)}, \tag{6.55}$$

$$A_3 := n^{\alpha^{\Theta,\mathcal{E}}_{\phi}} e^{\kappa^{\Theta,\mathcal{E}}_{\phi} n}, \tag{6.56}$$

$$A_4 := B^{\Theta,\mathcal{E}}_{s|\phi}, \tag{6.57}$$

$$A_5 := n^{\Delta^{\Theta,\mathcal{E}}_{s|\phi}}, \tag{6.58}$$

$$A_6 := \sum_{i=1}^{M} \frac{w(n) z_i^n}{Q^{\Theta}_{\phi,\mathcal{E}}(z_i, w)}, \tag{6.59}$$

$$A_7 := B_\phi^{\Theta,\mathcal{E}}, \tag{6.60}$$

and

$$A_8 := n^{\Delta_\phi^{\Theta,\mathcal{E}}}, \tag{6.61}$$

then Equation 6.53 becomes:

$$\rho_{[n_1,n_2]}^{\Theta,\mathcal{E}}(s|\phi) \sim A_1 A_2 \left[\frac{\sum_{n=n_1}^{n_2} A_3 \left(1 + \frac{A_4}{A_5}\right) A_6}{\sum_{n=n_1}^{n_2} A_3 \left(1 + \frac{A_7}{A_8}\right) A_6} \right]. \tag{6.62}$$

Equation 6.62 can be manipulated as follows:

$$\begin{aligned}
\rho_{[n_1,n_2]}^{\Theta,\mathcal{E}}(s|\phi) &\sim A_1 A_2 \left[\frac{\sum_{n=n_1}^{n_2} A_3 A_6}{\sum_{n=n_1}^{n_2} A_3 \left(1 + \frac{A_7}{A_8}\right) A_6} + \frac{A_4 \sum_{n=n_1}^{n_2} \frac{A_3}{A_5} A_6}{\sum_{n=n_1}^{n_2} A_3 \left(1 + \frac{A_7}{A_8}\right) A_6} \right] \\
&= A_1 A_2 \left[\left[1 + A_7 \frac{\sum_{n=n_1}^{n_2} \frac{A_3}{A_8} A_6}{\sum_{n=n_1}^{n_2} A_3 A_6} \right]^{-1} + \left[\frac{\sum_{n=n_1}^{n_2} A_3 \left(1 + \frac{A_7}{A_8}\right) A_6}{A_4 \sum_{n=n_1}^{n_2} \frac{A_3}{A_5} A_6} \right]^{-1} \right] \\
&= A_1 A_2 \left[\left[1 + A_7 \frac{\sum_{n=n_1}^{n_2} \frac{G_n}{A_8}}{\sum_{n=n_1}^{n_2} G_n} \right]^{-1} + \left[\frac{\sum_{n=n_1}^{n_2} G_n \left(1 + \frac{A_7}{A_8}\right)}{A_4 \sum_{n=n_1}^{n_2} \frac{G_n}{A_5}} \right]^{-1} \right] \\
&= A_1 A_2 \left[\left[1 + A_7 \frac{\sum_{n=n_1}^{n_2} \frac{G_n}{A_8}}{\sum_{n=n_1}^{n_2} G_n} \right]^{-1} + \frac{A_4 \sum_{n=n_1}^{n_2} \frac{G_n}{A_5}}{\sum_{n=n_1}^{n_2} G_n} \left[1 + A_7 \frac{\sum_{n=n_1}^{n_2} \frac{G_n}{A_8}}{\sum_{n=n_1}^{n_2} G_n} \right]^{-1} \right] \\
&= A_1 A_2 \left[1 + A_7 \frac{\sum_{n=n_1}^{n_2} \frac{G_n}{A_8}}{\sum_{n=n_1}^{n_2} G_n} \right]^{-1} \left[1 + \frac{A_4 \sum_{n=n_1}^{n_2} \frac{G_n}{A_5}}{\sum_{n=n_1}^{n_2} G_n} \right],
\end{aligned}$$

where $G_n = A_3 A_6 = n^{\alpha_\phi^{\Theta,\mathcal{E}}} e^{\kappa_\phi^{\Theta,\mathcal{E}} n} \sum_{i=1}^{M} \frac{w(n) z_i^n}{Q_{\phi,\mathcal{E}}^{\Theta}(z_i, w)}$.

If we substitute the original terms back into this equation, we obtain:

$$\begin{aligned}
\rho_{[n_1,n_2]}^{\Theta,\mathcal{E}}(s|\phi) &\sim \frac{Z_{n_1}^{\Theta,\mathcal{E}}(s|\phi)}{Z_{n_1}^{\Theta,\mathcal{E}}(\phi)} \frac{\left(1 + \frac{B_\phi^{\Theta,\mathcal{E}}}{n_1^{\Delta_\phi^{\Theta,\mathcal{E}}}}\right)}{\left(1 + \frac{B_{s|\phi}^{\Theta,\mathcal{E}}}{n_1^{\Delta_{s|\phi}^{\Theta,\mathcal{E}}}}\right)} \left[1 + B_\phi^{\Theta,\mathcal{E}} \frac{\sum_{n=n_1}^{n_2} \frac{G_n}{n^{\Delta_\phi^{\Theta,\mathcal{E}}}}}{\sum_{n=n_1}^{n_2} G_n}\right]^{-1} \left[1 + B_{s|\phi}^{\Theta,\mathcal{E}} \frac{\sum_{n=n_1}^{n_2} \frac{G_n}{n^{\Delta_{s|\phi}^{\Theta,\mathcal{E}}}}}{\sum_{n=n_1}^{n_2} G_n}\right] \\
&= \rho_{n_1}^{\Theta,\mathcal{E}}(s|\phi) \frac{\left(1 + \frac{B_\phi^{\Theta,\mathcal{E}}}{n_1^{\Delta_\phi^{\Theta,\mathcal{E}}}}\right)}{\left(1 + \frac{B_{s|\phi}^{\Theta,\mathcal{E}}}{n_1^{\Delta_{s|\phi}^{\Theta,\mathcal{E}}}}\right)} \left[1 + B_\phi^{\Theta,\mathcal{E}} \frac{\sum_{n=n_1}^{n_2} \frac{G_n}{n^{\Delta_\phi^{\Theta,\mathcal{E}}}}}{\sum_{n=n_1}^{n_2} G_n}\right]^{-1} \left[1 + B_{s|\phi}^{\Theta,\mathcal{E}} \frac{\sum_{n=n_1}^{n_2} \frac{G_n}{n^{\Delta_{s|\phi}^{\Theta,\mathcal{E}}}}}{\sum_{n=n_1}^{n_2} G_n}\right] \\
&\approx \rho_{n_1}^{\Theta,\mathcal{E}}(s|\phi) \text{ as } n_1 \to \infty.
\end{aligned}$$

Substituting in the scaling form for $\rho_{n_1}^{\Theta,\mathcal{E}}(s|\phi) = \frac{Z_{n_1}^{\Theta,\mathcal{E}}(s|\phi)}{Z_{n_1}^{\Theta,\mathcal{E}}(\phi)}$, up to first order $\rho_{n_1}^{\Theta,\mathcal{E}}(s|\phi)$ is approximately

$$\rho_{[n_1,n_2]}^{\Theta,\mathcal{E}}(s|\phi) \approx \frac{A_{s|\phi}^{\Theta,\mathcal{E}}}{A_\phi^{\Theta,\mathcal{E}}} + C n_1^{-D} \tag{6.63}$$

for some C and D.

Through the same arguments as above and the assumptions that $\alpha_{K|s,\phi}^{\Theta,\mathcal{E}} = \alpha_{s|\phi}^{\Theta,\mathcal{E}}$ and $\kappa_{K|s,\phi}^{\Theta,\mathcal{E}} = \kappa_{s|\phi}^{\Theta,\mathcal{E}}$, the grouped-$[n_1, n_2]$ probability of getting knot type K after strand passage

$$\rho_{[n_1,n_2]}^{\Theta,\mathcal{E}}(\phi \to K) := \frac{\sum_{n=n_1}^{n_2} \left[Z_n^{\Theta,\mathcal{E}}(K|\phi, s) \sum_{i=1}^{M} \frac{w(n) z_i^n}{Q_{\phi,\mathcal{E}}^{\Theta}(z_i, w)}\right]}{\sum_{n=n_1}^{n_2} \left[Z_n^{\Theta,\mathcal{E}}(s|\phi) \sum_{i=1}^{M} \frac{w(n) z_i^n}{Q_{\phi,\mathcal{E}}^{\Theta}(z_i, w)}\right]} \approx \rho_{n_1}^{\Theta,\mathcal{E}}(\phi \to K), \tag{6.64}$$

where up to first order $\rho_{[n_1,n_2]}^{\Theta,\mathcal{E}}(\phi \to K)$ is approximately

$$\rho_{[n_1,n_2]}^{\Theta,\mathcal{E}}(\phi \to K) \approx \frac{A_{K|s,\phi}^{\Theta,\mathcal{E}}}{A_{s|\phi}^{\Theta,\mathcal{E}}} + C' n_1^{-D'}, \tag{6.65}$$

for some C' and D' which depend on K.

The consequence of these results is that we can now use all the data from Θ-SAPs whose lengths are between n_1 and n_2 in the estimate corresponding to the length n_1. The width of the intervals $(n_2 - n_1)$ are determined similarly to the procedure described in Section 6.4; a test for independence is used to find values $n_1, n_2, \ldots, n_m$ which are essentially independent; then for $i = 1, \ldots, m-1$, the grouped n probabilities $\rho^{\Theta,\mathcal{E}}_{[n_i, n_{i+1}-2]}(s|\phi)$ and $\rho^{\Theta,\mathcal{E}}_{[n_i, n_{i+1}-2]}(\phi \to K)$ are used to estimate $\rho^{\Theta,\mathcal{E}}_{n_i}(s|\phi)$ and $\rho^{\Theta,\mathcal{E}}_{n_i}(\phi \to K)$.

6.6 Chapter Summary

Analyzing data coming from a CMC can be complicated. Although data coming from CMCs are (sometimes highly) correlated, the grouped-n method described in [61] for analyzing strand passage data allows one to consider all data coming from an interval of polygon lengths rather than throwing the majority of it away. Even though we are able to retain this data, there becomes a point where the error of the estimate relative to the estimate itself is intolerable and can possibly cause misleading inferences. Because of the complex nature of these problems, it is difficult to find an optimal solution. The methods presented in this chapter provide tools that minimize the chance of making misleading estimates. These techniques will be put into practice in the next two chapters: Chapter 7 will test the I-Pivot Algorithm developed in [64] and the I-Θ-BFACF algorithm developed here for consistency with other research, and Chapter 8 will present new results from simulations of multiple replications of the I-Θ-BFACF algorithm according to various salt concentrations.

Chapter 7

Algorithm Testing and Consistency

The purpose of this chapter is to check the consistency of the independently programmed I-Pivot Algorithm and the I-Θ-BFACF Algorithm. The consistency of these algorithms was checked by comparing some preliminary simulation results with those obtained from other sources. It is important to note that all error bars that appear in this work reflect a 95% confidence interval on the parameter of interest. All of the analysis performed in this Chapter as well as Chapter 8 was done using code written in C and R.

7.1 Θ-BFACF Algorithm

In order to ensure that the copy of the new I-Θ-BFACF algorithm developed here was functioning correctly, a short simulation was conducted with the energy turned off (*i.e.* $A/k_BT = 0$ and $v = 0$); this is equivalent to the good solvent case. One of the easiest consistency checks that can be made for the I-Θ-BFACF algorithm is comparing the average length of Θ-SAPs (at equilibrium) in a particular chain; thus, the average equilibrium length for each chain was estimated in the simulation and was compared to the averages presented in [61].

It should be noted that the code for the Θ-BFACF algorithm without Metropolis sampling was obtained from M. Szafron, the author of [60] and [61]. This code was modified to include Metropolis sampling based on SAP energy to obtain the I-Θ-BFACF algorithm and is programmed in C. Using the updating scheme for SAP energy described in Section 5.6, the run time for a simulation is linear with n. For the average polygon lengths considered in Chapters 7 and 8 (around 200-300 in the highest chain), and for a CMC consisting of 10 chains, it typically takes around 9 hours to run 1 billion time steps on the Bugaboo cluster of Compute Canada's *Westgrid* computing network, which uses a Intel Xeon E5430 quad-core processor running at 2.66 GHz.

7.1.1 Simulation Details

A CMC implementation of the I-Θ-BFACF algorithm was run for 1 billion time steps with $A/k_BT = 0$ and $v = 0$; the CMC consisted of 10 different chains with z-values $Z_1 = 0.197, Z_2 = 0.2, Z_3 = 0.203, Z_4 = 0.205, Z_5 = 0.207, Z_6 = 0.2090, Z_7 = 0.21, Z_8 = 0.2105, Z_9 = 0.211, Z_{10} = 0.2115$, where Z_i corresponds to the fugacity z in the i^{th} chain. The fugacities Z_3 to Z_{10} were chosen to be the same as those used for chains 1 to 8 in [61]; Z_1 and Z_2 were selected so that the corresponding chains yield a very small average length; this generally speeds up the time to convergence to equilibrium for the CMC. The parameter $q = 2$ was used, swapping between a randomly selected pair of adjacent chains was attempted every 5 time steps, and samples were taken every 1000 time steps. The choices of q, swap frequency and sample rate are also the same choices that were made in [61]. The average Θ-SAP length in chains 3 to 10 will be compared to those obtained in [61].

7.1.2 Warm-up Analysis

Because the average equilibrium Θ-SAP length in chain 10 is the function that has the highest variance out of all parameters of interest, a warm-up analysis was performed using this function to estimate τ_{exp}. This warm-up analysis, which is shown in Figure 7.1, shows that the trend of the 1 to j column averages is relatively steady after about 150 million time steps. Going from right to left in Figure 7.1, the trend of the j to n column averages stops fluctuating wildly after about 200 million time steps. Thus, an estimate of τ_{exp} for this CMC system is 200 million time steps.

7.1.3 Estimating τ_{int}

Since the correlation between datapoints will also be higher in chain 10 than any other chain, the batch means procedure was used on the data coming after the first 200 million time steps from chain 10 to estimate $2\tau_{\text{int}}$ for all the chains. This procedure yielded an estimate of $2\tau_{\text{int}} = 45$ million time steps. Therefore, two datapoints are considered essentially independent if they are separated by 45 million time steps or more.

7.1.4 Estimating Average Θ-SAP Length

Table 7.1 shows the estimates of the average Θ-SAP length of polygons coming from chains 3 to 10 of the above CMC implementation of the Θ-BFACF algorithm. These estimates assume that $\tau_{\text{exp}} = 200$ million time steps and $\tau_{\text{int}} = 45$ million time steps; denote the estimate for the average

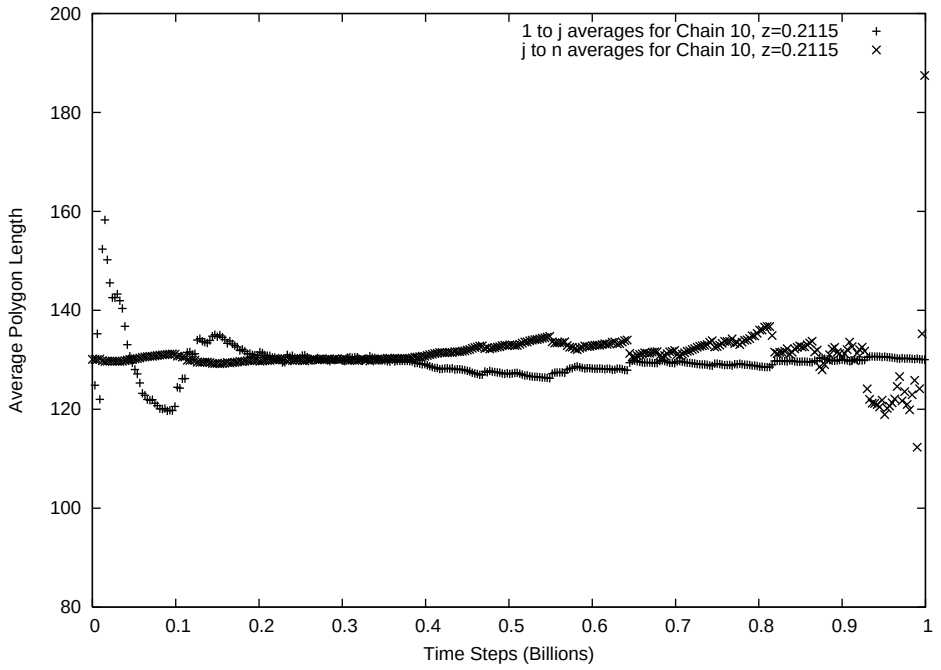

Figure 7.1: Warmup analysis for chain 10 of the Θ-BFACF algorithm, where $q = 2$ and $z = 0.2115$.

length in chain i by $\langle n_{z_i}(\mathcal{P}^{\Theta}(\phi)\rangle$. Column 1 of Table 7.1 contains the number of each chain. Column two contains the z-values (*i.e.* fugacity) for each chain. Column 3 shows the estimates for $\langle n_{z_i}(\mathcal{P}^{\Theta}(\phi))\rangle$ along with their corresponding 95% margins of error. Column 4 shows the results for $\langle n_{z_i}(\mathcal{P}^{\Theta}(\phi))\rangle$ obtained in [61], along with the corresponding 95% margins of error. Table 7.1 shows that the estimates obtained here are statistically comparable to the results in [61]. It should be noted that the 95% error bars obtained here are smaller than those obtained in [61], even though the estimates presented here are based on less independent data. A possible reason for this observation might be due to the fact that the CMC implementation of the Θ-BFACF algorithm in [61] had additional chains which yielded much larger average Θ-SAP lengths compared to the chains considered here. Because of the definition of the CMC, a state from a higher chain (with a large average polygon length) can possibly be swapped into a lower chain (with a smaller average polygon length). If this transition happens over a short amount of time steps, then the length of the state that was originally in the higher chain may not be representative of the distribution of the smaller chain when it gets swapped there. This correlation between chains can inflate confidence intervals when estimating quantities such as average length.

Chain i	Fugacity z_i	$\langle n_{z_i}(\mathcal{P}^{\Theta}(\phi))\rangle$	Estimate from [61]
3	0.203	34.9 ± 0.1	34.9 ± 1.6
4	0.205	40.1 ± 0.1	40.0 ± 2.0
5	0.207	48.4 ± 0.2	48.5 ± 2.6
6	0.209	64.3 ± 0.5	64.7 ± 3.8
7	0.21	79.5 ± 0.7	80.0 ± 4.9
8	0.2105	90.8 ± 1.2	91.6 ± 5.8
9	0.211	106.8 ± 2.2	108.0 ± 7.0
10	0.2115	130.1 ± 4.9	132.9 ± 8.8

Table 7.1: Average Θ-SAP length for different chains compared with those obtained in [61].

7.2 Pivot Algorithm with Energy

Several simulations of the I-Pivot Algorithm were run for 50 million time steps. Simulations had varied polygon lengths (n=200, 300 and 400), ζ values ranging from 0.1 to 10 and the value $A/k_BT = 0.01$. Each simulation consisted of a Markov chain with a unique set of parameter values (n, A/k_BT, ζ); composite Markov chains were not involved. The results obtained here will be qualitatively compared with the results from the DNA experiments of Shaw and Wang [53] and the simulations of Tesi *et al.* that show DNA chain knotting probability increases with salt concentration.

The Pivot Algorithm and I-Pivot Algorithm were both programmed from scratch in C. The energy of a SAP must be calculated at each time step; thus, the run time of the simulation takes $O(n^2)$ time, where n is the length of the polygon. It is interesting to note that when ζ is small, SAPs tend to be less compact; this causes an increase in the chance of a pivot move to result in a polygon that is still self avoiding. This results in more successful pivots on average, which causes the program to run longer. Thus, the run time actually increases as ζ decreases.

7.2.1 Estimating τ_{exp}

The observable quantity from $\mathcal{H}'$ (i.e. the set of observable functions of interest) over all the simulations that was judged to have the highest variance was the mean square radius of gyration for $\zeta = 0.1$, $n = 400$ and $A/k_BT = 0.01$. Therefore, a warm-up analysis conducted on this quantity should provide a sufficient upper bound for τ_{exp} for all of the simulations. This warm-up analysis is presented in Figure 7.2.

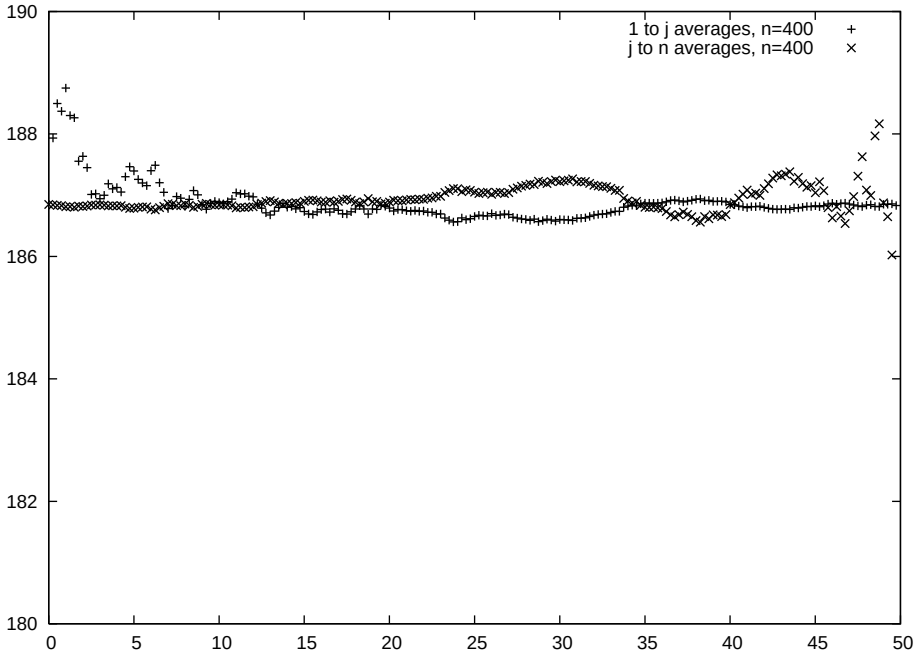

Figure 7.2: Warm-up analysis for the I-pivot algorithm with parameters $\zeta = 0.1$, $n = 400$ and $A/k_BT = 0.01$

Because the trend of the first j column averages and the last j column averages shown in Figure 7.2 dissipates around 5 million time steps, Section 3.2.3 tells us we can use 5 million time steps as an estimate for an upper bound for τ_{exp}. Although it may appear that these column averages have not dissipated after 5 million time steps, when one considers the scale of the y-axis in Figure 7.2, one can see that these fluctuations are less than 1% of the total average.

The data corresponding to these first 5 million time steps are *burned*, *i.e.* not used in the final analysis. Thus, the analysis of quantities of interest will only focus on time steps greater than τ_{exp}.

7.2.2 Comparison of Mean Square Radius of Gyration

The batch means procedure described in Section 3.2.6 was performed on the data corresponding to time steps greater than τ_{exp} (i.e. equilibrium data). This blocking procedure for the square radius of gyration data produced an estimate for $2\tau_{\text{int}}$ of 65000 time steps. Thus, we say that two batches of square radius of gyration data are *essentially independent* if they are separated by 65000 time steps or more.

Estimates for the mean square radius of gyration were calculated for all of the simulations assuming that τ_{exp} is 5 million time steps and $2\tau_{\text{int}}$ is 65000 time steps. Table 7.2 shows a complete list of the results obtained here. Figure 7.3 compares these results with those obtained by

Tesi *et al.* in [64]. It is worth noting that the datapoints from [64] are only approximate, as they did not provide a table for their results. These approximate values were interpolated from [64, Fig. 4] and these values are also presented in Table 7.2.

n	A	ζ	Estimate for Mean $\mathcal{R}^2$	Best Guess for Mean $\mathcal{R}^2$ from [64]
200	0.01	0.10	59.7 ± 0.2	57
200	0.01	0.32	45.7 ± 0.1	44
200	0.01	0.56	38.9 ± 0.1	37.2
200	0.01	1.00	34.9 ± 0.2	34.3
200	0.01	3.16	32.7 ± 0.2	32.6
200	0.01	10.00	32.7 ± 0.2	32.9
300	0.01	0.10	116.5 ± 0.4	111
300	0.01	0.32	78.1 ± 0.2	74.5
300	0.01	0.56	62.9 ± 0.3	60
300	0.01	1.00	54.2 ±0.3	52.8
300	0.01	3.16	49.7 ±0.3	49.8
300	0.01	10.00	49.5 ± 0.3	49.5
400	0.01	0.10	186.5 ± 0.6	177
400	0.01	0.32	114.3 ± 0.4	107
400	0.01	0.56	89.0 ± 0.4	83.8
400	0.01	1.00	74.2 ± 0.4	71.5
400	0.01	3.16	67.5 ± 0.5	66.9
400	0.01	10.00	66.4 ± 0.5	67.1

Table 7.2: 95% confidence intervals for the mean square radius of gyration for different values of n and ζ, and interpolated values from Figure 4 in [64].

Figure 7.3 reveals that some of the estimates for smaller values of ζ and larger lengths obtained here are not statistically similar with those obtained in [64]. However, they differ only by a small order of magnitude; also, the datapoints obtained here follow the trend of what one would expect to see in this model (i.e. a decrease in the mean square radius of gyration as ζ increases). It should also be noted that as ζ becomes larger, the estimates agree quite well with those obtained in [64].

The difference between the corresponding datapoints can possibly be explained by the fact that the data generated in [64] was via a simulation with a length of 2.5 million time steps. In this work

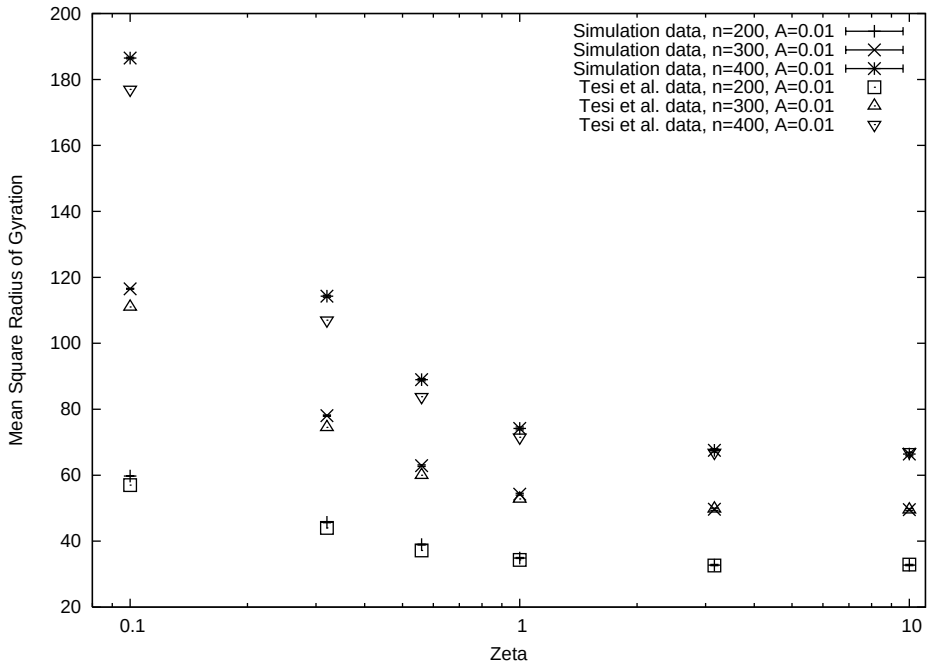

Figure 7.3: Estimates for the mean square radius of gyration with $A/k_BT = 0.01$ compared with Tesi *et al.* in [64].

it was estimated that the exponential autocorrelation time for the system was 5 million time steps; although this is a conservative estimate, it opens up the possibility that some of the data obtained in [64] may not have been from the equilibrium distribution. It is also possible that there were differences due to greater round-off error due to the limitations in the availability and speed of high precision computers in the 1980s and 1990s. The authors of [64] were contacted to request their data and programmed versions of their algorithms; however, both their data and programs were no longer available.

7.2.3 Mean Number of Contacts

The estimated mean number of contacts (presented in Table 7.3 and illustrated in Figure 7.4) was computed for all of the simulations under the assumption that $\tau_{\text{exp}} = 5$ million time steps. The batch means procedure yielded an estimate of $2\tau_{\text{int}} = 42500$ time steps. As shown in Figure 7.3, SAPs corresponding to higher salt concentrations are on average more compact. Thus, it is expected that the mean number of contacts will increase as the salt concentration gets larger. Figure 7.4 shows that this is indeed the case.

n	A	ζ	Estimate for Mean Contacts
200	0.01	0.10	54.2 ± 0.1
200	0.01	0.32	63.3 ± 0.1
200	0.01	0.56	70.0 ± 0.2
200	0.01	1.00	75.9 ± 0.3
200	0.01	3.16	80.4 ± 0.3
200	0.01	10.00	80.5 ± 0.3
300	0.01	0.10	77.8 ± 0.2
300	0.01	0.32	92.1 ± 0.2
300	0.01	0.56	103.0 ± 0.3
300	0.01	1.00	113.3 ± 0.4
300	0.01	3.16	120.4 ± 0.5
300	0.01	10.00	121.4 ± 0.6
400	0.01	0.10	102.2 ± 0.2
400	0.01	0.32	121.1 ± 0.3
400	0.01	0.56	135.8 ± 0.4
400	0.01	1.00	150.0 ± 0.6
400	0.01	3.16	160.7 ± 0.8
400	0.01	10.00	162.5 ± 0.8

Table 7.3: 95% confidence intervals for the mean number of contacts for different values of n and ζ.

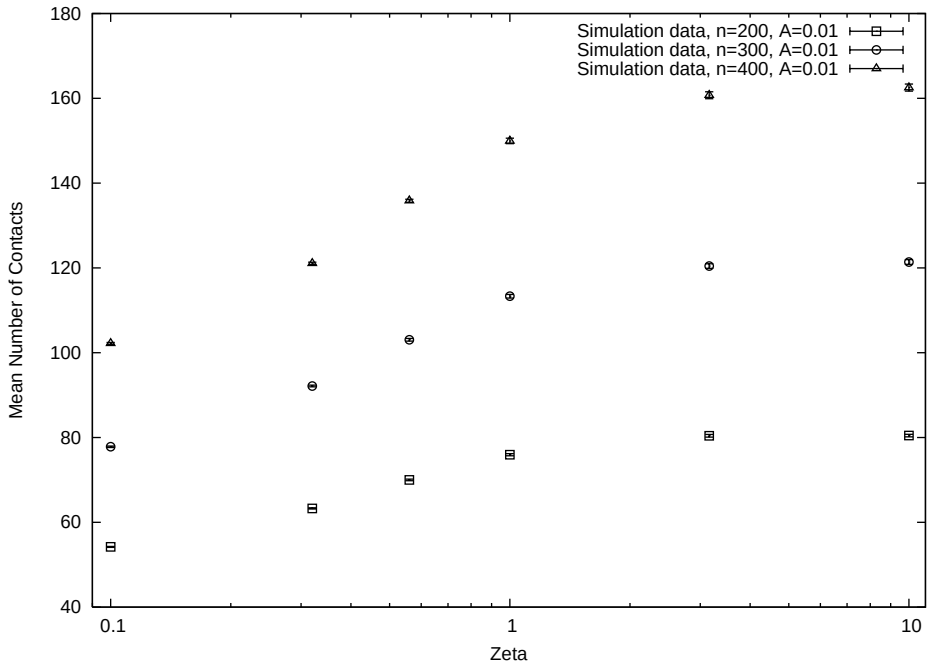

Figure 7.4: Estimates for the mean number of contacts for different values of n and ζ.

7.2.4 Knotting Probability

One quantity that is very reflective as to whether or not the model is having its intended effect is the probability of a SAP being knotted. DNA experiments from Shaw and Wang [53] show that the probability of a DNA chain being knotted increases with salt concentration.

Similar to the mean square radius of gyration and mean number of contacts, the probability of knotting was estimated for $n = 200$, 300, and 400 and ζ ranging from 0.1 to 10. The same estimate of $\tau_{\text{exp}} = 5$ million time steps was used as the burntime. The batch means procedure yielded an estimate for $2\tau_{\text{int}}$ of 62000 time steps. The results for these simulations are presented in Table 7.4 and illustrated in Figure 7.5. Figure 7.5 shows a clear increase in the probability of knotting with increasing length and salt concentration.

In order to compare the results obtained here with the results of Shaw and Wang in [53], the values of ζ used in the simulations were converted to concentrations using the relation in Equation 2.31. These conversions are shown in Table 7.5. The next issue that needs to be addressed is determining how many base pairs are modelled by one lattice edge. Under normal physiological conditions, the effective helical diameter of DNA has been measured to be 5nm; this length corresponds to the span of approximately 15 base pairs of double helix DNA [71]. If we consider this effective helical diameter to represent one lattice unit (*i.e.* the excluded volume in the simple cubic

n	A	ζ	Estimate for Knotting Probability
200	0.01	0.10	0.0003 ± 0.0001
200	0.01	0.32	0.0015 ± 0.0003
200	0.01	0.56	0.0027 ± 0.0012
200	0.01	1.00	0.0052 ± 0.0017
200	0.01	3.16	0.0071 ± 0.0018
200	0.01	10.00	0.0068 ± 0.0013
300	0.01	0.10	0.0006 ± 0.0004
300	0.01	0.32	0.0017 ± 0.0003
300	0.01	0.56	0.0068 ± 0.0027
300	0.01	1.00	0.0097 ± 0.0017
300	0.01	3.16	0.0117 ± 0.0018
300	0.01	10.00	0.0141 ± 0.0023
400	0.01	0.10	0.0007 ± 0.0003
400	0.01	0.32	0.0023 ± 0.0005
400	0.01	0.56	0.0075 ± 0.0012
400	0.01	1.00	0.0160 ± 0.0026
400	0.01	3.16	0.0232 ± 0.0042
400	0.01	10.00	0.0262 ± 0.0045

Table 7.4: 95% confidence intervals for the probability of knotting for different values of n and ζ.

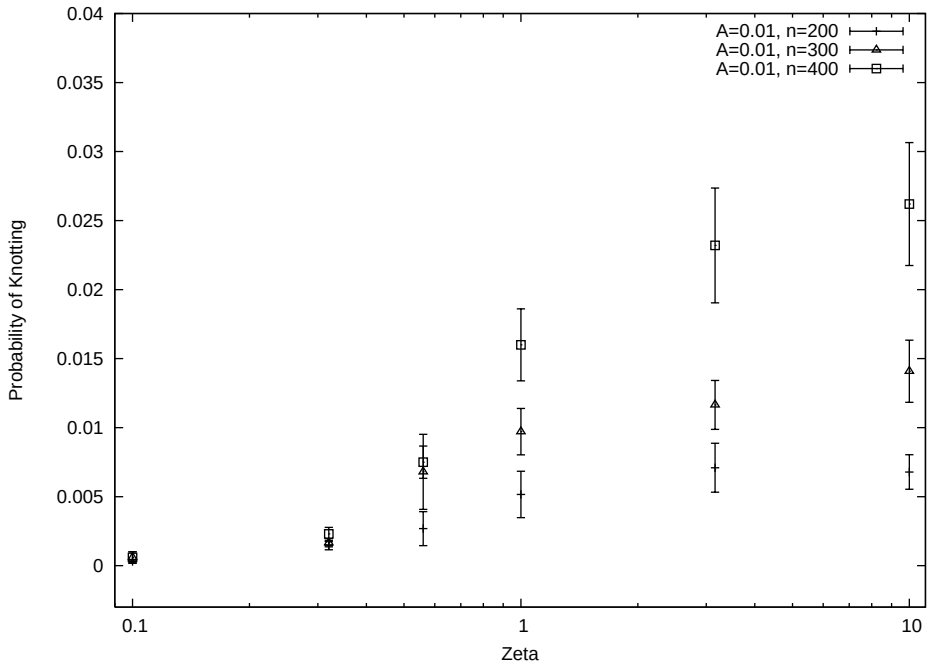

Figure 7.5: Estimates for the probability of knotting for different values of n and ζ.

lattice), then a SAP with 400 edges models circular DNA with approximately 6 kbp. Although the helical diameter of DNA changes with salt concentration [51], this estimate of 15 bp/edge provides a starting point for comparing results from SAPs to results from DNA.

ζ	NaCl Concentration (mol/L)
0.1	0.00093
0.32	0.0092
0.56	0.03
1	0.093
3.16	0.93
10	9.3

Table 7.5: Conversion of ζ values to concentrations of NaCl in mol/L.

Figure 7.6 shows the comparisons of knotting probabilities between the data from [53] for DNA chains with 8.6 kilo base pairs (kbp) and the simulation data for $n = 400$. Note that the estimates corresponding to the data from [53] that are plotted were interpolated from the graph that the authors provided in [53], and thus are only approximate estimates; these approximate estimates are listed in Table 7.6.

It is important to note that the comparison in Figure 7.6 is only intended to show the qualitative trends of the data, and not to compare exact results. This comparison suggests that the knotting probabilities obtained here for polygons of length 400 are qualitiatively comparable to the knotting probability for 8.6 kbp DNA obtained in [53].

NaCl Concentration (mol/L)	Estimate for Knotting Probability
0.05	0.0075 ± 0.001
0.1	0.012 ± 0.0010
0.2	0.015 ± 0.0015
0.5	0.023 ± 0.0020
1	0.028 ± 0.0050
2	0.030 ± 0.0040

Table 7.6: Approximate estimates for the knotting probabilities of 8.6 kbp DNA chains as a function of NaCl concentration presented by Shaw and Wang in [53].

7.3 Chapter Summary

All of the results presented here are consistent with what one would expect from the model; more specifically, as the salt concentration increases, we see a decrease in mean square radius of gyration, increase in the mean number of contacts, and an increase in knotting probability. The comparison of the knotting probability results obtained here with those obtained by Shaw and Wang in [53] indicates that the energy model is a good way to model DNA in solution. One can also note that there becomes a point where increasing ζ no longer provides a significant change in these quantities. This can be viewed physically as the solution being 'saturated' with salt. Mathematically, the reason for this is because the term involving ζ, namely $Ae^{-\zeta r_{ij}}/r_{ij}$, becomes negligible as ζ becomes large; thus, at some point increasing ζ will provide no significant difference in the results. Now that we have established that the energy model is a good way to model DNA in solution, we now turn to the case of strand passages in SAPs with a fixed structure.

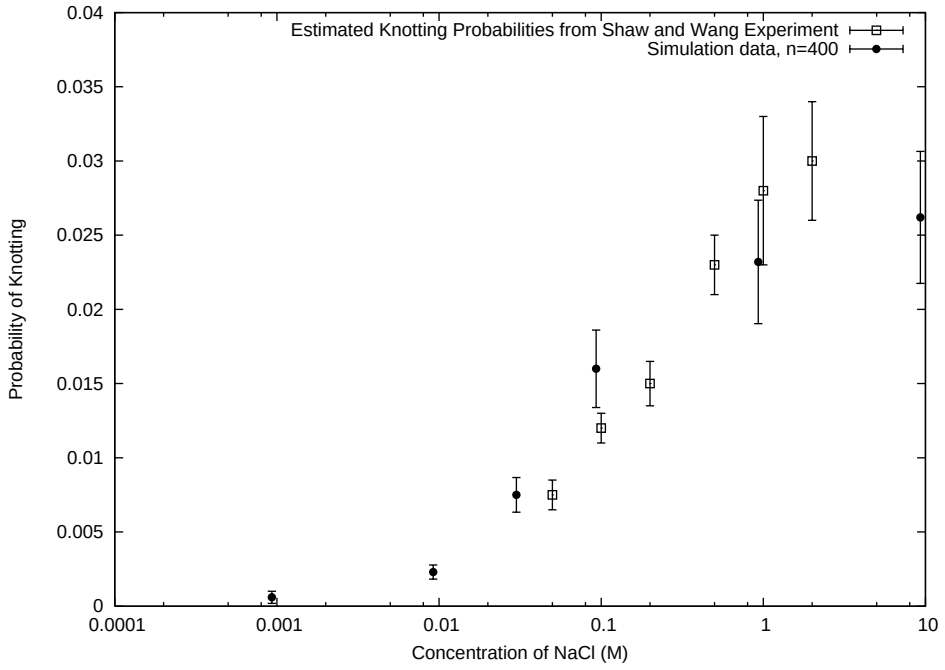

Figure 7.6: Estimates for the probability of knotting for polygons with length $n = 400$ plotted as a function of NaCl concentration and compared to the knotting probabiltiies obtained in [53] for DNA chains with 8.6 kbp.

Chapter 8

Results from the I-Θ-BFACF Algorithm

The following chapter presents the results obtained from several CMC implementations of the new I-Θ-BFACF algorithm on unknotted Θ-SAPs; these implementations covered a wide range of ζ values (*i.e.* salt concentrations). The average equilibrium polygon length was estimated for each chain from the simulations corresponding to each value of ζ. These estimates were used to provide a rough estimate of the critical value $z_c^{\mathcal{E}}(\phi)$ for the different ζ-values. Results for the mean square radius of gyration are presented for $\zeta = 0.1$ and $\zeta = 10$ to show the effect that added salt has on the average volume a Θ-SAP occupies.

In order to address Problem 2, the strand passage and knot transition probabilities $\rho_n^{\Theta,\mathcal{E}}(s|\phi)$, $\rho_n^{\Theta,\mathcal{E}}(\phi \to \phi)$, and $\rho_n^{\Theta,\mathcal{E}}(\phi \to 3_1^+)$ were estimated for each salt concentration using grouped-n estimation. However, it was not always possible to get a good estimate for the limiting knot transition probabilities $\rho^{\Theta,\mathcal{E}}(\phi \to \phi)$ and $\rho^{\Theta,\mathcal{E}}(\phi \to 3_1^+)$ because a goodness of fit test failed using the estimated region of reliable data.

8.1 Simulation Details

A series of CMC implementations of the I-Θ-BFACF Algorithm were carried out for 10 different values of ζ, namely $\zeta = \{0.1, 0.2, 0.56, 0.8, 1, 1.5, 2.2, 3.16, 6, 10\}$, where $A/k_BT = 0.01$ and $v = -0.26$ for all simulations. From this point on when A and v are specified, it will be assumed that $A/k_BT = 0.01$ and $v = -0.26$. For each value of ζ, 10 independent replications were run for 40 billion time steps each, where each replication was started in a different starting state.

In an attempt to have these starting states be 'relatively far apart' (refer to Section 3.2.4), preliminary simulations of 10 billion time steps were run for each value of ζ. The state of the chain after i billion time steps was selected to be the starting state for the i^{th} replication. Each simulation was a CMC consisting of 10 chains, where the fugacities for each chain depended on the value of ζ being considered. A complete list of fugacities for each chain and ζ value is shown in Table 8.1.

Swapping was attempted between a pair of adjacent chains every 5 time steps, and samples were taken every 10,000 time steps. The initial random number seeds for each simulation were selected using a separate random number generator scheme in MATLAB or R. The algorithms used are written in C and run on the Bugaboo cluster of Compute Canada's *Westgrid* network, which uses a Intel Xeon E5430 quad-core processor running at 2.66 GHz. It usually takes 9-12 hours to run a simulation for 1 billion time steps; however, since the algorithm takes order $O(n)$ time to run where n is the length of the polygon in the chain, it can take 48 hours or longer to do a billion time steps when one or more chains contain large polygons. The code used to perform the analysis in this chapter was written in C or R.

Finding the proper distribution of z-values for each ζ is not an easy task; as mentioned in Section 5.5, the first step is to find a z value which converges to equilibrium quite rapidly (*i.e.* has a smaller average length at equilibrium). From this value, the z-value is slightly increased until a z-value is found which yields an average equilibrium length of around 200 to 400. This z-value is chosen to be the fugacity of the highest chain. A z-value that obtains a fairly small average equilibrium length (around 30) is chosen to be the fugacity for the first chain. Select the fugacities for the rest of the chains so that they are equally spaced between the highest and lowest z-value. At this point, an algorithm described in [60, page 97] is used to distribute the remaining z-values in a way that produces nearly optimal swap rates between adjacent chains.

ζ \ Chain	1	2	3	4	5	6	7	8	9	10
0.1	0.2050	0.2079	0.2103	0.2122	0.2137	0.2149	0.2159	0.2167	0.2174	0.2179
0.2	0.1975	0.2001	0.2027	0.2049	0.2067	0.2081	0.2092	0.2100	0.2104	0.2107
0.56	0.1947	0.1960	0.1972	0.1981	0.1992	0.2000	0.2008	0.2013	0.2018	0.2021
0.8	0.1884	0.1913	0.1937	0.1956	0.1972	0.1983	0.1992	0.1998	0.2003	0.2006
1	0.1884	0.1912	0.1935	0.1953	0.1968	0.1978	0.1986	0.1992	0.1996	0.1999
1.5	0.1880	0.1904	0.1924	0.1939	0.1952	0.1962	0.1970	0.1976	0.1981	0.1985
2.2	0.1884	0.1907	0.1926	0.1942	0.1954	0.1963	0.1970	0.1975	0.1979	0.1982
3.16	0.1884	0.1906	0.1923	0.1937	0.1949	0.1958	0.1965	0.1972	0.1976	0.1979
6	0.1892	0.1909	0.1923	0.1935	0.1946	0.1955	0.1962	0.1967	0.1971	0.1975
10	0.1892	0.1908	0.1922	0.1935	0.1945	0.1954	0.1961	0.1967	0.1971	0.1975

Table 8.1: A list of fugacities for each chain and value of ζ.

8.2 Using Potential Scale Reduction to Estimate $\tau_{\mathbf{exp}}$

To determine when the simulations had reached equilibrium, the estimated potential scale reduction was calculated using C over the 10 replications for each value of ζ. Recall from Section 3.2.4 that a series of simulations are deemed to have converged to their equilibrium distribution at the point where the estimated potential scale reduction is consistently below 1.05. This time can be used as an estimate for τ_{exp}; define this estimate to be $\hat{\tau}_{\text{exp}}(\zeta)$ for the CMC simulations with a given ζ. Graphs of all the estimated potential scale reductions are shown in Appendix A; the corresponding estimates for $\hat{\tau}_{\text{exp}}(\zeta)$ are in Table 8.2.

ζ	$\hat{\tau}_{\text{exp}}(\zeta)$
0.1	12.0×10^9
0.2	0.4×10^9
0.56	0.1×10^9
0.8	0.2×10^9
1	0.5×10^9
1.5	0.1×10^9
2.2	0.2×10^9
3.16	0.1×10^9
6	0.1×10^9
10	0.1×10^9

Table 8.2: The estimates $\hat{\tau}_{\text{exp}}(\zeta)$ for each value of ζ.

In Section 3.2.2, it was mentioned that if the estimate for $\tau_{\text{exp}}(\zeta)$ is less than 5% of the total run time, then the data from the first $\tau_{\text{exp}}(\zeta)$ time steps can be included in the final analysis. For $\zeta = 0.1$, the estimate for $\tau_{\text{exp}}(\zeta)$ (12 billion time steps) is 30% of the total run time (40 billion time steps); therefore the data coming from the first 12 billion time steps of those simulations is not used in the final analysis. However, for all other values of ζ, the estimate for $\tau_{\text{exp}}(\zeta)$ is 1.25% or less of the total run time; hence, all of the data from these simulations are included in the final analysis.

8.3 Using Batch Means to Estimate τ_{int}

The batch means procedure outlined in Section 3.2.6 is used in R to calculate an estimate of τ_{int} for every independent replication. The data in this procedure is only the data coming from time steps after $\tau_{\text{exp}}(\zeta)$. Define $\tau_{\text{int}}(\zeta, i)$ to be the estimate of τ_{int} for the i^{th} replication of a particular value of ζ, and define $\tau_{\text{int}}(\zeta)$ to be the maximum of $\tau_{\text{int}}(\zeta, i)$ over all 10 replications; *i.e.*

$$\tau_{\text{int}}(\zeta) = \max_{1 \leq i \leq 10} \tau_{\text{int}}(\zeta, i). \tag{8.1}$$

The estimates for $2 \times \tau_{\text{int}}(\zeta)$ are shown in Table 8.3. The estimates of $2 \times \tau_{\text{int}}(\zeta, i)$ for every replication is shown in Appendix B.

ζ	$2 \times \hat{\tau}_{\text{int}}(\zeta)$
0.1	1.73×10^9
0.2	0.51×10^9
0.56	0.17×10^9
0.8	0.18×10^9
1	0.30×10^9
1.5	0.12×10^9
2.2	0.09×10^9
3.16	0.13×10^9
6	0.11×10^9
10	0.08×10^9

Table 8.3: The estimates of $2 \times \hat{\tau}_{\text{int}}(\zeta)$ for each value of ζ.

8.4 Mean Square Radius of Gyration

As a simple check to see whether the I-Θ-BFACF Algorithm was consistent with the results from the I-Pivot Algorithm, the mean square radius of gyration was calculated for the I-Θ-BFACF Algorithm simulations with $\zeta = 0.1$ and $\zeta = 10$ and compared to those obtained in the I-Pivot Algorithm for the same values of ζ. Because polygons in the I-Θ-BFACF Algorithm are forced to have the Θ-structure, the mean square radius of gyration for Θ-SAPs of length n will be slightly different than the mean square radius of gyration for unrooted SAPs of the same length.

In order to compare the SAPs from the I-Pivot Algorithm (where the knot type is allowed to vary) with the unknotted SAPs coming from the I-Θ-BFACF algorithm, estimates for the mean square radius of gyration were recalculated for the I-Pivot Algorithm data using only the SAPs that were unknotted. It was found that these new estimates of the mean square radius of gyration corresponding only to unknotted SAPs did not differ significantly from the estimates when all SAPs were considered; this is because a large majority (more than 97%, see Figure 7.5) of the SAPs in the I-Pivot Algorithm were unknots.

Estimates for the mean square radius of gyration corresponding to the I-Θ-BFACF algorithm data were calculated assuming the values of $\hat{\tau}_{\text{exp}}(\zeta)$ and $2 \times \hat{\tau}_{\text{int}}(\zeta)$ for $\zeta = 0.1$ and 10 presented in Tables 8.2 and 8.3. Figure 8.1 shows how differently the mean square radius of gyration grows with n when $\zeta = 0.1$ compared to when $\zeta = 10$.

Table 8.4 shows a comparison of the estimates of the mean square radius of gyration for $n = 200, 300, 400$ and $\zeta = 0.1, 10$ from the I-Pivot Algorithm and the I-Θ-BFACF Algorithm.

ζ	n	Mean $\mathcal{R}^2$ (I-Θ-BFACF)	Mean $\mathcal{R}^2$ (I-Pivot)
0.1	200	56.4 ± 0.7	59.7 ± 0.2
0.1	300	109.8 ± 1.5	116.5 ± 0.4
0.1	400	175.8 ± 2.8	186.5 ± 0.6
10	200	30.7 ± 0.02	32.7 ± 0.2
10	300	47.3 ± 0.05	49.6 ± 0.3
10	400	64.1 ± 0.1	66.9 ± 0.5

Table 8.4: Comparison of the Mean Square Radius of Gyration estimates from the I-Pivot Algorithm (unknotted SAPs only) and the I-Θ-BFACF algorithm along with the estimated 95% margins of error.

In the comparison of mean square radius of gyration estimates in Table 8.4, one will notice that the estimates from the I-Θ-BFACF Algorithm are consistently lower than the estimates from the I-Pivot Algorithm. Recall that the Θ-structure represents two segments of a SAP being pulled close together. Because Θ-SAPs are forced to contain this fixed structure, one would expect that the mean square radius of gyration for Θ-SAPs would be slightly smaller than in the unconstrained case.

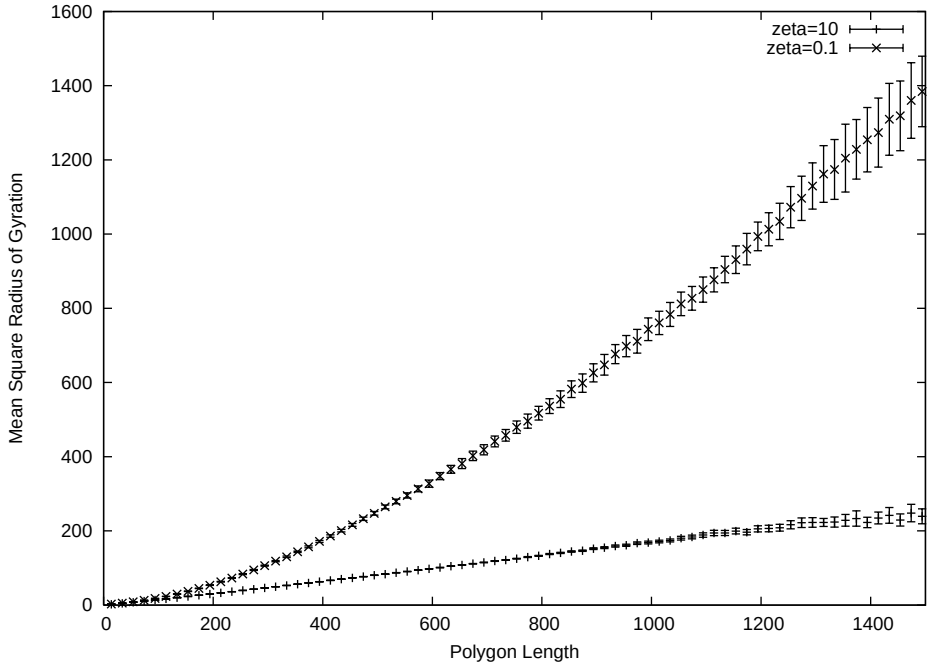

Figure 8.1: How the mean square radius of gyration of a Θ-SAP grows with length for $\zeta = 10$ and $\zeta = 0.1$.

8.5 Average Polygon Length

Tables 8.5 and 8.6 present the estimates for the average length of a polygon in each chain of the CMC for each value of ζ. Refer to Table 8.1 for the z-value that corresponds to a particular chain and ζ.

The bolded value in Table 8.6 corresponds to the estimate of the average length in chain 10 for the simulations with $\zeta = 0.1$. The large margin of error for this estimate (roughly 50% of the point estimate) is reflective of the autocorrelation issues described in Section 5.5 that occur due to very large polygons appearing in the highest chains of simulations with small values of ζ. This increased variability leads to a much larger estimated standard error, and hence a larger estimated margin of error.

8.6 Estimating the critical value $\mathbf{z_c^{\Theta,\mathcal{E}}}(\phi)$ for Different Values of ζ

If the Θ-BFACF algorithm is used to sample unknots in the good solvent case, then the *critical z-value* $z_c(\phi) = e^{-\kappa_\phi}$. In this case, as z approaches $z_c(\phi)$, the plot of $1/z$ versus $1/\bar{n}_z$ (where $\bar{n}_z$ is defined to be the average polygon length of a chain with fugacity z) should become linear as

ζ \ Chain	1	2	3	4	5
0.1	38.0 (0.2)	43.3 (0.2)	49.3 (0.2)	56.0 (0.3)	63.2 (0.3)
0.2	34.0 (0.1)	38.0 (0.1)	43.8 (0.2)	51.1 (0.2)	60.7 (0.3)
0.56	44.9 (0.2)	50.3 (0.2)	57.3 (0.3)	64.4 (0.3)	76.1 (0.4)
0.8	34.4 (0.1)	40.6 (0.2)	49.0 (0.2)	60.7 (0.3)	77.1 (0.4)
1	36.1 (0.1)	43.0 (0.2)	52.6 (0.2)	66.0 (0.3)	86.1 (0.4)
1.5	37.5 (0.1)	44.1 (0.2)	52.6 (0.2)	63.7 (0.3)	78.6 (0.4)
2.2	39.5 (0.2)	47.3 (0.2)	57.6 (0.3)	71.7 (0.3)	91.0 (0.5)
3.16	40.1 (0.2)	47.4 (0.2)	56.9 (0.3)	69.5 (0.3)	86.2 (0.4)
6	42.6 (0.2)	49.0 (0.2)	57.0 (0.3)	67.6 (0.3)	81.4 (0.4)
10	42.6 (0.2)	48.9 (0.2)	56.7 (0.3)	66.8 (0.3)	80.2 (0.4)

Table 8.5: Average lengths for chains 1 to 5 and each ζ value. Numbers in parentheses are the estimated 95% margins of error.

ζ \ Chain	6	7	8	9	10
0.1	70.5 (0.4)	79.9 (0.5)	92.7 (0.9)	115.4 (5.6)	**212.0 (111.2)**
0.2	73.1 (0.3)	91.3 (0.6)	118.9 (1.5)	157.8 (5.0)	215.5 (19.8)
0.56	91.7 (0.5)	112.1 (0.6)	138.6 (0.9)	172.1 (1.6)	213.2 (3.5)
0.8	100.3 (0.5)	133.3 (0.7)	178.3 (1.2)	237.6 (2.6)	318.6 (8.1)
1	114.4 (0.6)	154.3 (0.9)	211.2 (1.6)	290.5 (3.8)	395.7 (11.3)
1.5	98.7 (0.5)	125.0 (0.7)	162.0 (1.0)	211.0 (1.7)	276.6 (4.2)
2.2	117.3 (0.6)	152.1 (0.9)	197.6 (1.3)	257.2 (2.5)	332.8 (5.5)
3.16	108.8 (0.6)	138.6 (0.8)	178.6 (1.1)	231.0 (2.0)	298.5 (4.4)
6	99.2 (0.5)	121.9 (0.6)	150.6 (0.9)	184.6 (1.3)	228.8 (2.3)
10	97.1 (0.5)	119.0 (0.6)	147.9 (0.9)	182.3 (1.2)	228.9 (2.3)

Table 8.6: Average lengths for chains 6 to 10 and each ζ value. Numbers in parentheses are the estimated 95% margins of error.

$1/z \to 1/z_c(\phi)$, where the x-intercept of this line corresponds to $1/z_c(\phi)$ [61]. In order to determine whether a similar trend holds in the case of the I-Θ-BFACF algorithm, a plot of $1/z$ versus $1/\bar{n}_{z,\zeta,A,\phi}$ (where $\bar{n}_{z,\zeta,A,K}$ is defined to be the average polygon length of a chain with fugacity z and choice of A and ζ and knot type K) was created for different values of ζ. If these plots become linear as $1/z$ decreases to $1/z_c^{\Theta,\mathcal{E}}(\phi)$, then a regression line can be computed, where the x-intercept of this line can provide an estimate for $z_c^{\Theta,\mathcal{E}}(\phi)$.

Figure 8.2 shows that in the simulations where $\zeta = 0.1$, there was no apparent linear trend in the plot of $1/z$ versus $1/\bar{n}_{z,\zeta,A,\phi}$ as $1/z$ decreases. This could be because the values of $1/z$ considered are not close enough to $1/z_c^{\Theta,\mathcal{E}}(\phi)$ in order to approach this linear trend; another reason for this lack of linearity might be due to the large error in the estimate for the average length in chain 10.

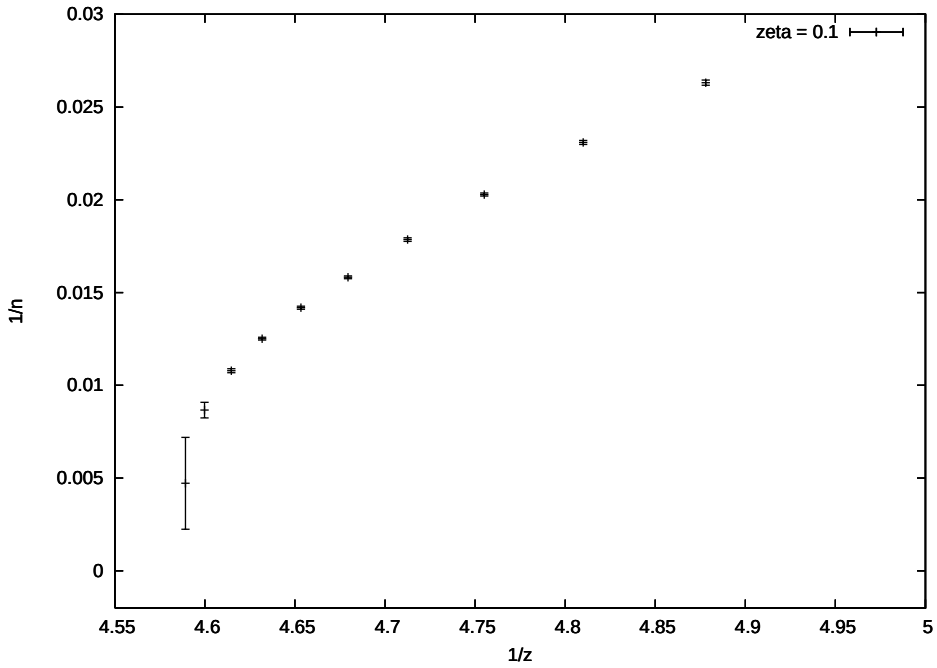

Figure 8.2: A plot of $1/z$ versus $1/\bar{n}_{z,\zeta,A,\phi}$ for $\zeta = 0.1$.

The graph in Figure 8.2 highlights some of the problems that occurred in the simulations with $\zeta = 0.1$; the autocorrelation time related to these simulations is very large when compared to the simulations for other values of ζ. The massive error bars in the estimate for the average length of chain 10 in this simulation raises the concern that the z value corresponding to that chain might be larger than the critical z-value, thus making it a divergent chain.

In the simulations for all other values of ζ, a similar plot of $1/z$ versus $1/\bar{n}_{z,\zeta,A,\phi}$ was graphed to look for a linear trend as $1/z$ decreases. Figure 8.3 shows the plot of $1/z$ versus $1/\bar{n}_{z,\zeta,A,\phi}$ where

$\zeta = 1$. The plot for this value of ζ is displayed because the estimate for average equilibrium polygon length in chain 10 for $\zeta = 1$ was higher than the estimated average equilibrium polygon length in all the chains from all the other simulations. One can immediately see that this plot is closer to linearity as $1/z$ decreases than the graph in Figure 8.2. However, there is still a slight non-linear trend in this data. A close-up of this graph is shown in Figure 8.4.

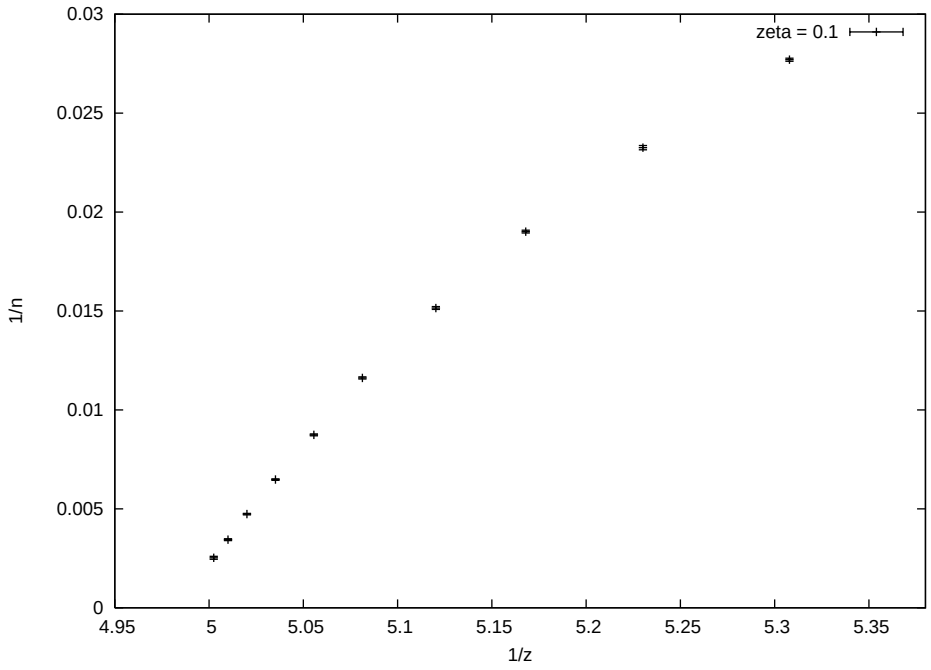

Figure 8.3: A plot of $1/z$ versus $1/\bar{n}_{z,\zeta,A,\phi}$ for $\zeta = 1$.

The graph shown in Figure 8.4 indicates that there is still evidence of non-linearity in the plot of the points being considered. Thus, a regression line between the points corresponding to the few largest z-values may not be appropriate. Instead, to get a rough estimate of $z_c^{\Theta,\mathcal{E}}(\phi)$ for each value of ζ, it was decided to plot the line that goes through the two points in the graph of $1/z$ versus $1/\bar{n}_{z,\zeta,A,\phi}$ corresponding to the last two chains (*i.e.* the two largest z-values) for that particular value of ζ. The x-intercept of this line is then a rough estimate for the value of $1/z_c^{\Theta,\mathcal{E}}(\phi)$. These estimates are listed in Table 8.7 for each value of ζ.

It should be noted that these estimates for $z_c^{\Theta,\mathcal{E}}(\phi)$ are likely upper bounds for the actual values of $z_c^{\Theta,\mathcal{E}}(\phi)$, so one should be careful when choosing z-values close to these estimates. However, if one is able to find a z-value that converges to a higher average equilibrium polygon length (for some $\mathcal{E}$), one could get a better estimate for $z_c^{\Theta,\mathcal{E}}(\phi)$.

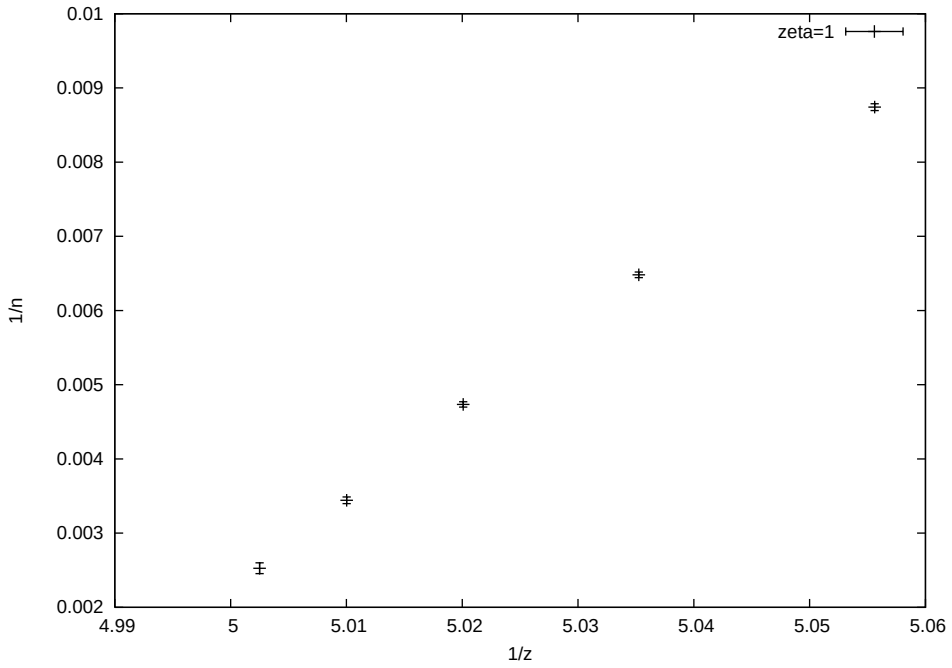

Figure 8.4: A close-up of the plot of $1/z$ versus $1/\bar{n}_{z,\zeta,A,\phi}$ for $\zeta = 1$.

ζ	Estimate for $z_c^{\Theta,\mathcal{E}}(\phi)$
0.1	0.2185003
0.2	0.2115248
0.56	0.2033660
0.8	0.2014852
1	0.2007331
1.5	0.1997976
2.2	0.1992275
3.16	0.1989336
6	0.1991883
10	0.1990805

Table 8.7: Estimates for the critical z-value $z_c^{\Theta,\mathcal{E}}(\phi)$ for each value of ζ.

8.7 Limiting Successful Strand Passage Probabilities

The probability of a successful strand passage was calculated using ratio estimation for each observed polygon length n. Using grouped-n estimation, this data will be used to numerically explore the existence of limiting successful strand passage probabilities. Recall in Section 6.3 that not all of the data from the simulation can be used; eventually there is a polygon length $N_{\max}(*)$ where the data becomes "unreliable". The next section will provide a detailed example for how this region of reliable data is calculated.

8.7.1 Reliable Data Example

The following section provides an example of the reliable data procedure outlined in Section 6.3. Figure 8.5 shows the fixed-n estimates of $\rho_n^{\Theta,\mathcal{E}}(s|\phi)$ (as defined in Equation 6.43) for n up to 1500 from the simulation where $\zeta = 1$.

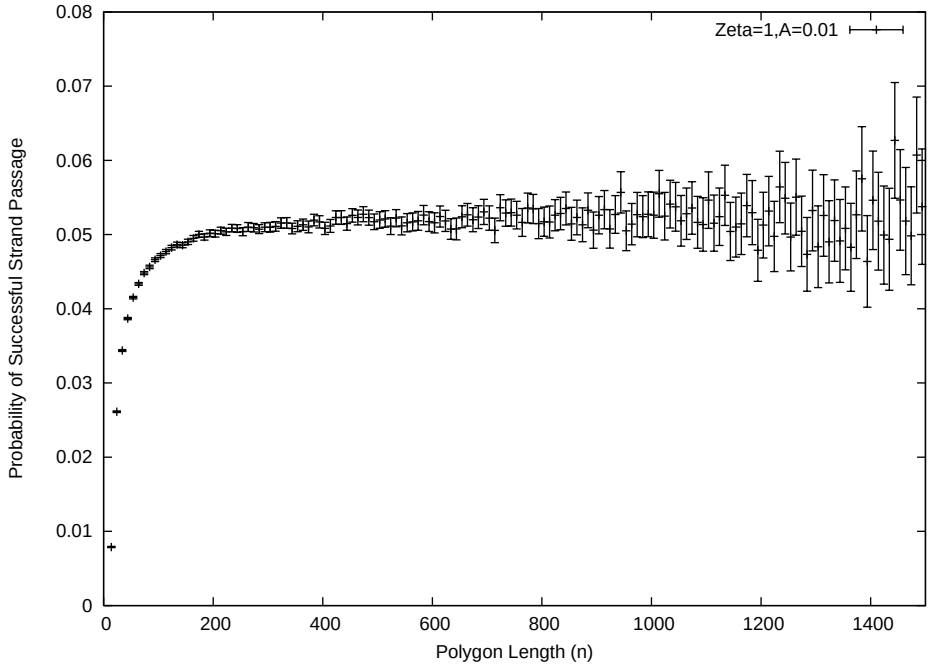

Figure 8.5: Estimates for the probability of successful strand passage when $\zeta = 1$.

One can observe that the error bars in Figure 8.5 are increasing significantly as the polygon length gets to $n = 1000$ and beyond. In order to obtain an estimate for the maximum polygon length for which the estimates of $\rho_n^{\Theta,\mathcal{E}}(s|\phi)$ are reliable (define this length to be $\hat{N}_{\max}(s|\phi,\mathcal{E})$), the relative standard error of $\rho_n^{\Theta,\mathcal{E}}(s|\phi)$ needs to be plotted for each of the ten independent replications.

Recall from Equation 6.30 that the relative standard error of a point estimate is the standard error of the point estimate divided by the point estimate itself. The relative error of $\rho_n^{\Theta,\mathcal{E}}(s|\phi)$ is plotted for each replication in Figure 8.6.

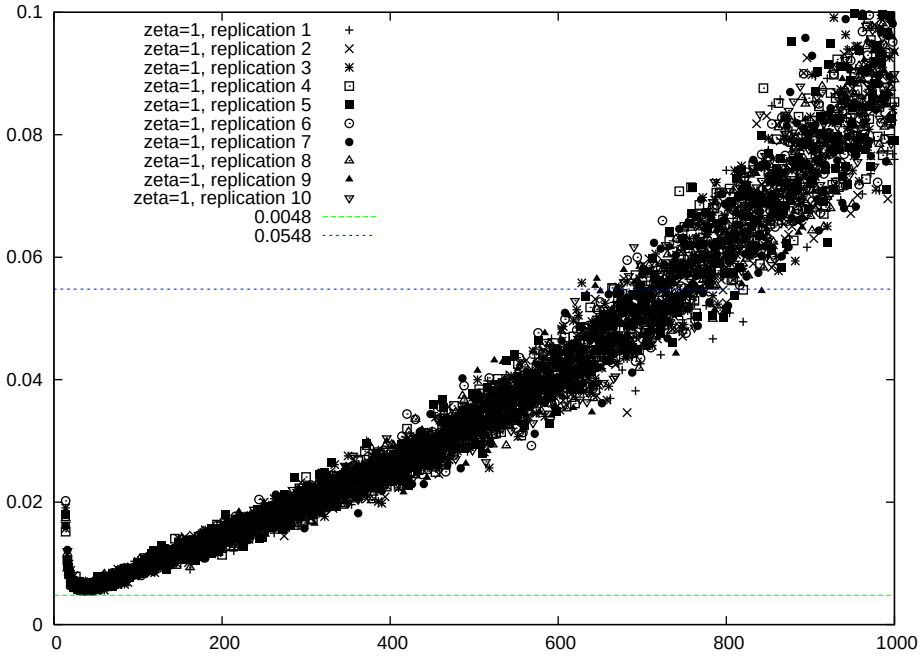

Figure 8.6: Plot of the relative standard error to determine $\hat{N}_{\max}(s|\phi,\zeta,A)$

The lower dashed line in Figure 8.6 is the minimum relative standard error obtained over all 10 replications; recall from Section 6.3 that this corresponds to $\min_r \hat{\delta}^{(r)}(s|\phi,\zeta,A)$. This is estimated to be 0.0048. The upper dashed line in Figure 8.6 is the cutoff point $\epsilon_* = \min_r(\hat{\delta}^{(r)}(*)+c)$. Recall that c, the maximum tolerated deviation from $\min_r \hat{\delta}^{(r)}(*)$, is chosen to be 0.05. Thus, $\epsilon_* = 0.0548$. If $\hat{\eta}^{(r)}(s|\phi,\zeta,A)$ represents the value of n that achieves the minimum relative standard error of 0.0048, then $\hat{N}_{\max}(s|\phi,\zeta,A)$ is the first value of n after $\hat{\eta}^{(r)}(s|\phi,\zeta,A)$ such that the relative standard error exceeds ϵ_*. For this particular case, $\hat{N}_{\max}(s|\phi,\zeta,A) = 628$. Thus, for $\zeta = 1$, only the data corresponding to polygon lengths less than 628 is considered "reliable".

Using only the estimates corresponding to polygon lengths less than $\hat{N}_{\max}(s|\phi,\zeta,A)$ for each value of ζ, the batch means procedure described in Section 3.2.6 was used to determine how far apart values of n need to be for the estimates of $\rho_n^{\Theta,\mathcal{E}}(s|\phi)$ to be "essentially independent". Recall from Section 6.5 that using the grouped-n analysis technique, one can use the data coming from all of the polygon lengths in between these essentially independent datapoints and treat it like it came from the first datapoint. A complete list of these independent batch sizes, as well as estimates of

$\hat{N}_{\max}(s|\phi,\zeta,A)$, for each value of ζ is shown in Table 8.8.

ζ	$\hat{N}_{\max}(s\|\phi,\zeta,A)$	Independent Batch Size
0.1	230	40
0.2	420	58
0.56	454	72
0.8	552	88
1	628	100
1.5	536	86
2.2	568	92
3.16	558	90
6	490	78
10	468	74

Table 8.8: Estimates for the amount of reliable data $\hat{N}_{\max}(s|\phi,\zeta,A)$ and the independent batch size corresponding to $\hat{N}_{\max}(s|\phi,\zeta,A)$ for different ζ.

8.7.2 Estimates for the Limiting Successful Strand Passage Probability

Using the independent batch sizes and reliable data specified in Table 8.8 for each value of ζ, the probabilities of a successful strand passage were calculated using the grouped-n method (programmed in C). Figure 8.7 shows the estimates relating to this grouping procedure for $\zeta = 1$.

In Section 6.5, recall that as $n_1 \to \infty$, the grouped-n probabilities are expected to scale like $B + Cn^{-D} + O(n^{-1})$, where B, C and D all depend on the knot type and energy being used. The value of B is the limiting successful strand passage probability and is of particular interest. However, for some values of ζ used here it appears that the corrections contained in the $O(n^{-1})$ term for smaller values of n are significant enough that it is difficult to get a good fit to the form $B + Cn^{-D}$. Including higher order correction terms of the form En^{D-1} or Fn^{-1} can give a better fit. Even with these extra terms, the first datapoint (corresponding to length 14) is not at all representative of the asymptotic scaling form, and is not included in the fit.

Non-linear least squares regression in the *R* programming environment was used to fit the data to one of the models listed above. Table 8.9 shows the results for the fits for all values of ζ, including a goodness of fit test (described next) for each fit.

If $\{(X_1, Y_1), \ldots, (X_m, Y_m)\}$ represents the m essentially independent datapoints that are being

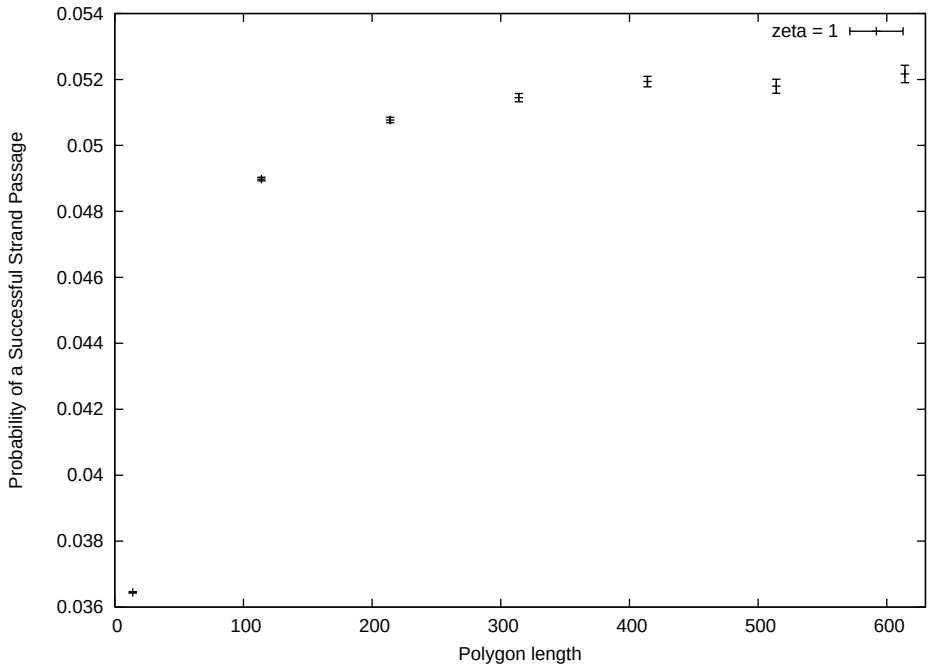

Figure 8.7: Grouped-n estimates for the probability of successful strand passage when $\zeta = 1$.

fit, $\{\hat{\sigma}_1^2, \ldots, \hat{\sigma}_m^2\}$ represents the estimated variances of the estimates $Y_1, \ldots, Y_m$, and $\{(X_1, Z_1), \ldots, (X_n, Z_m)\}$ are the expected datapoints of the fit at points $X_1, \ldots, X_m$, then the goodness of fit test statistic used here is

$$T = \sum_{i=1}^{m} \frac{(Y_i - Z_i)^2}{\hat{\sigma}_i^2}. \tag{8.2}$$

If there are M parameters in the model being fitted, then there are $m - M$ degrees of freedom for the test. If Q is a random variable such that $Q \sim \chi^2_{m-M}$, then the p-value corresponding to this test statistic is $\Pr(Q \geq T)$. The null hypothesis for this test is that the model being fitted is appropriate. If the p-value is less than 0.05 for a particular fit, then there is reasonable evidence to suggest that the model being fitted may not be appropriate.

A graph of the fits obtained for $\zeta = 0.2, 0.8, 3.16, 10$ is shown in Figure 8.8. Note that there is little difference between the fits for $\zeta = 3.16$ and $\zeta = 10$; this suggests that the addition of salt at this point is no longer significant (*i.e.* the solution is saturated).

It should be noted that there is a systematic error that arises in estimating limiting probabilities due to the choice of a minimum polygon length for which the model is being fitted. In these fits this minimum polygon length is chosen to be the length corresponding to the second independent datapoint in the region of reliable data; this length is chosen because the first essentially independent

ζ	Regression Fit	$\hat{\rho}^{\Theta,\mathcal{E}}(s\|\phi)$ (S.E.)	G.O.F. Test Statistic	df	p-value
0.1	$0.057 + 2.369n^{-1} - 166.4n^{-2}$	0.057 (0.0006)	2.69	2	0.26
0.2	$0.057 + 0.311n^{-0.674} - 20.567n^{-1.674}$	0.057 (0.001)	0.63	3	0.89
0.56	$0.058 - 2.565n^{-1.397}$	0.058 (0.0001)	1.17	3	0.76
0.8	$0.055 - 0.460n^{-1}$	0.055 (< 0.0001)	3.42	4	0.49
1	$0.053 - 0.442n^{-1}$	0.053 (< 0.0001)	7.01	4	0.13
1.5	$0.050 - 0.382n^{-1}$	0.050 (< 0.0001)	4.14	3	0.25
2.2	$0.050 - 0.097n^{-0.645}$	0.050 (0.0003)	1.37	2	0.50
3.16	$0.048 - 0.379n^{-1}$	0.048 (< 0.0001)	5.22	4	0.26
6	$0.048 - 0.367n^{-1}$	0.048 (< 0.0001)	5.45	4	0.24
10	$0.048 - 0.180n^{-0.821} - 0.358n^{-1.821}$	0.048 (0.0008)	16.1	3	0.001

Table 8.9: The results of the fits for the successful strand passage probabilities, estimates of the limiting successful strand passage probabilities, and statistics pertaining to a goodness of fit test on each regression fit.

datapoint does not fit the asymptotic form of the limiting probability. However, due to the small amount of essentially independent datapoints in these fits, such a systematic error was not able to be estimated.

The bolded value in Table 8.9 indicates that the fit for $\zeta = 10$ failed the goodness of fit test. This is due to one datapoint (for $n = 458$) that deviates from the trend of the others. Because there are not very many datapoints, this deviation greatly affected the quality of the fit. One can notice in Table 8.9 that the limiting probability of a successful strand passage decreases as ζ increases; this is because higher salt concentrations yield more compact polygons on average, thus causing the vertices around the Θ-structure to be occupied more frequently. Refer to Figure 2.1 to see which vertices must be unoccupied for a successful strand passage to occur.

8.8 Limiting Knot Transition Probabilities

Table 8.10 presents the estimates for the amount of reliable data as well as the independent batch sizes for the estimates of $\hat{\rho}_n^{\Theta,\mathcal{E}}(\phi \to K)$ (as defined in Equation 6.44) for each $K \in \{\phi, 3_1\}$ and for each value of ζ. Recall from Section 1.2 that the trefoil knot (*i.e.* 3_1) is chiral, with these chiralities denoted by 3_1^+ and 3_1^-. Also recall from Section 2.1.1 that a strand passage in a class II Θ-SAP (*i.e.* a Θ^--SAP) switches a negative crossing to a positive crossing. Because the Θ-SAPs considered in

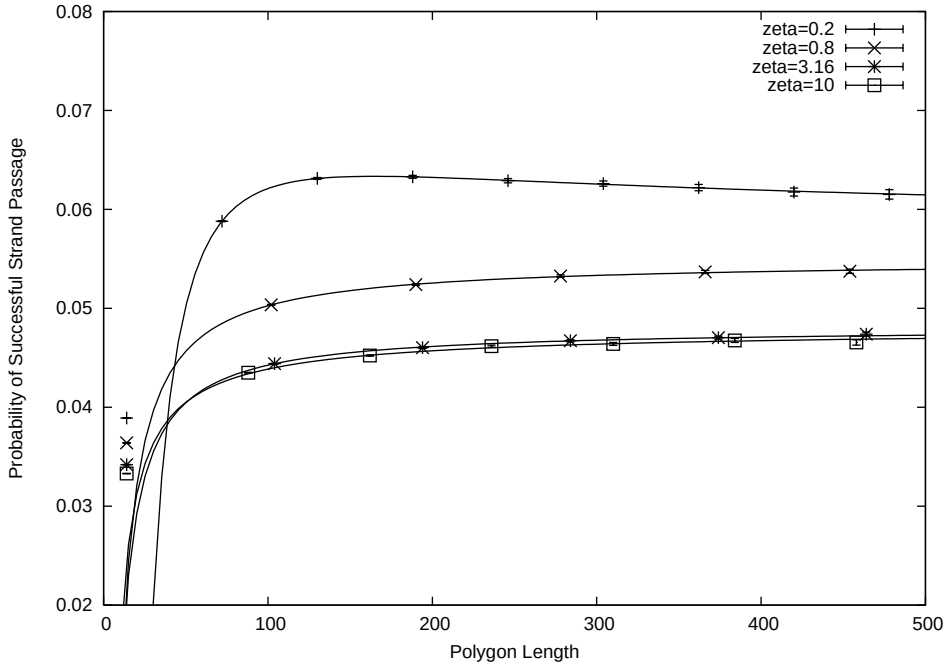

Figure 8.8: Grouped-n estimates of the probability of successful strand passage for $\zeta = 0.2, 0.8, 3.16, 10$ along with their asymptotic scaling form fits.

the simulations of the I-Θ-BFACF Algorithm are all Θ^--SAPs, this negative to positive crossing strand passage is the only type of strand passage that can occur. It has been reviewed by Soteros *et al.* in [56] that if a successful strand passage on an unknotted Θ^--SAP yields a trefoil knot, then this trefoil must be a 3_1^+ knot. Thus, we are only interested in estimating the knot transition probability $\hat{\rho}_n^{\Theta,\mathcal{E}}(\phi \to K)$ for each $K \in \{\phi, 3_1^+\}$.

Table 8.11 displays the frequencies with which different knot types were obtained after a successful strand passage from the simulation data for each ζ value. From these results it is clear that the majority of knotted after strand passage SAPs have the knot type 3_1^+.

Fits were attempted on the knot transition probabilities $\hat{\rho}_n^{\Theta,\mathcal{E}}(\phi \to \phi)$ and $\hat{\rho}_n^{\Theta,\mathcal{E}}(\phi \to 3_1^+)$ using grouped-n estimation in the region of reliable data in order to estimate the limiting knot transition probabilities $\hat{\rho}^{\Theta,\mathcal{E}}(\phi \to \phi)$ and $\hat{\rho}^{\Theta,\mathcal{E}}(\phi \to 3_1^+)$. The procedure used to generate these estimates is identical to the methods described in Section 8.7.

8.8.1 Unknot to Unknot

The results for the fits of $\hat{\rho}_n^{\Theta,\mathcal{E}}(\phi \to \phi)$ based on the regions of reliable data defined by $\hat{N}_{\max}(\phi \to \phi|\zeta, A)$ are presented in Table 8.12. Examples of these fits for 6 values of ζ are shown in Figure 8.9. One will notice from these results that as ζ increases, the limiting probability of a Θ-SAP remaining

ζ	$\hat{N}_{\max}(\phi \to \phi \vert \zeta, A)$	Batch Size	$\hat{N}_{\max}(\phi \to 3_1^+ \vert \zeta, A)$	Batch Size
0.1	290	44	94	20
0.2	550	64	142	30
0.56	634	116	180	38
0.8	914	164	234	48
1	1058	190	232	48
1.5	816	156	210	44
2.2	916	176	246	52
3.16	802	154	234	50
6	688	138	240	50
10	672	136	224	48

Table 8.10: Estimates for the amount of reliable data and the independent batch size corresponding to the limiting knot transition probabilities $\hat{\rho}_n^{\Theta,\mathcal{E}}(\phi \to \phi)$ and $\hat{\rho}_n^{\Theta,\mathcal{E}}(\phi \to 3_1^+)$.

ζ	ϕ	3_1^+	4_1	5 crossings or more
0.1	18335136	191989	3403	67
0.2	18499169	213984	4641	104
0.56	17464519	349614	14619	881
0.8	16365151	396880	22573	1876
1	16154016	448583	29677	2861
1.5	15240863	406245	26412	2486
2.2	15187658	461298	33871	4019
3.16	14952391	437468	30510	3403
6	14796328	400501	25436	2517
10	14761653	395220	25248	2837

Table 8.11: Observed counts of after-strand passage knot types from the I-Θ-BFACF simulations for each value of ζ.

an unknot after a successful strand passage decreases.

ζ	Fit	$\hat{\rho}^{\Theta,\mathcal{E}}(\phi \to \phi)$ (S.E.)	G.O.F. Test Statistic	df	p
0.1	$0.998 - 0.372n^{-0.589} + 14.090n^{-1.589}$	0.998 (0.002)	0.02	1	0.89
0.2	$0.983 - 17.4n^{-1.536} + 1261.7n^{-2.536}$	0.983 (0.0007)	1.60	4	0.81
0.56	$0.952 + 3.684n^{-1.166}$	0.952 (0.0015)	1.05	2	0.59
0.8	$0.937 + 2.798n^{-1}$	0.937 (0.0002)	0.37	3	0.95
1	$0.931 + 3.266n^{-1}$	0.931 (0.0003)	0.41	3	0.94
1.5	$0.904 + 0.515n^{-0.486}$	0.904 (0.022)	4.05	2	0.13
2.2	$0.880 + 0.338n^{-0.323}$	0.880 (0.014)	0.34	2	0.85
3.16	$0.901 + 0.744n^{-0.555}$	0.901 (0.0045)	0.31	2	0.86
6	$0.889 + 0.512n^{-0.432}$	0.889 (0.019)	0.76	1	0.38
10	$0.869 + 0.417n^{-0.333} - 0.358n^{-1.333}$	0.869 (0.012)	0.10	1	0.75

Table 8.12: The results of the fits for the limiting knot transition probabilities of going from $\phi \to \phi$, estimates for this limiting probability with standard error, and statistics pertaining to a goodness of fit test on each regression fit.

One can observe in Table 8.12 that for large values of ζ, the limiting knot transition probability $\hat{\rho}^{\Theta,\mathcal{E}}(\phi \to \phi)$ is about 0.9. This is a substantial decrease from the results in the good solvent case, where this limiting transition probability is estimated to be 0.97653 ± 0.00133 [61]. This tells us that when the salt concentration is high, it is more likely that an unknotted Θ-SAP will become a knot after a strand passage than in the good solvent case.

8.8.2 Unknot to Trefoil

Recall from Table 8.10 that the region of reliable data defined by $\hat{N}_{\max}(\phi \to 3_1^+|\zeta, A)$ is less than 250 for each value of ζ. In Figure 8.10 one can observe from the plots of $\hat{\rho}_n^{\Theta,\mathcal{E}}(\phi \to 3_1^+)$ (for five different values of ζ) that this knot transition probability has not yet reached its asymptotic form in the region of reliable data (*i.e.* for $n < 250$). There are also only four independent data points less than $\hat{N}_{\max}(\phi \to 3_1^+|\zeta, A)$ for each value of ζ. This can cause a problem because fitting such few datapoints for smaller values of n may not be very informative. Nevertheless, the results of the fits using the region of reliable data defined by $\hat{N}_{\max}(\phi \to 3_1^+|\zeta, A)$ are presented in Table 8.13.

Although several of the fits in Table 8.13 pass the goodness of fit test (at the 5% significance level), the standard error of the estimate for $\hat{\rho}^{\Theta,\mathcal{E}}(\phi \to 3_1^+)$ in most cases is quite large. If we

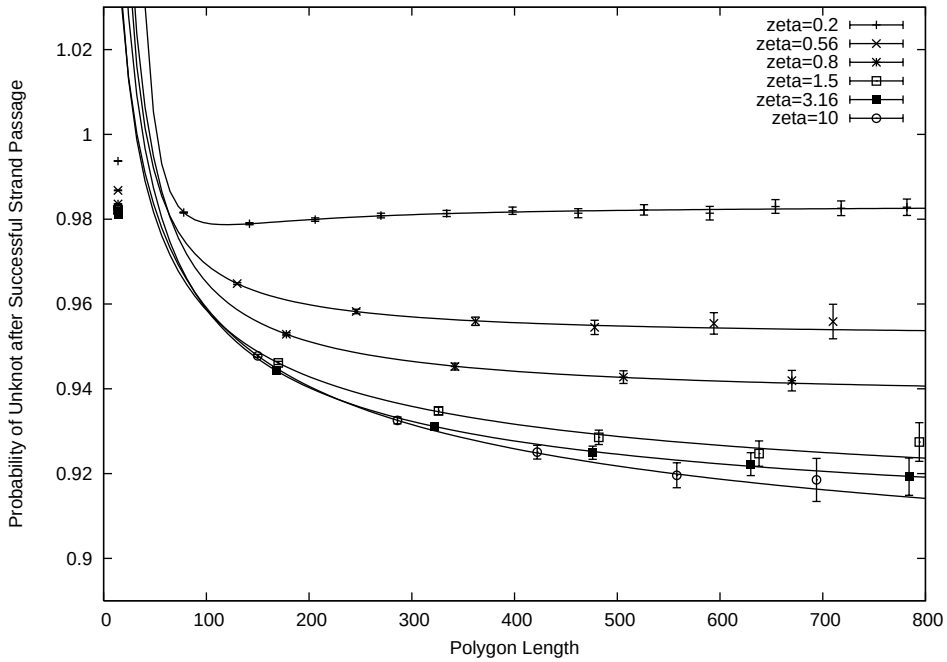

Figure 8.9: Grouped-n estimates of $\hat{\rho}_n^{\Theta,\mathcal{E}}(\phi \to \phi)$ for various choices of ζ along with their asymptotic scaling form fits.

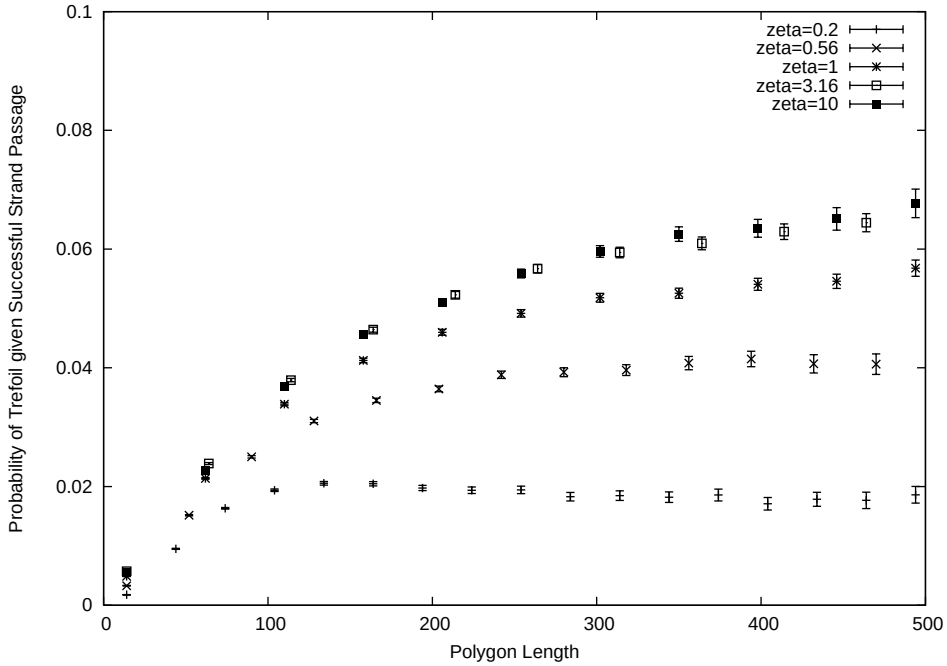

Figure 8.10: Grouped-n estimates of $\hat{\rho}_n^{\Theta,\mathcal{E}}(\phi \to 3_1^+)$ for various choices of ζ (without fits, due to not enough informative data).

ζ	Fit	$\hat{\rho}^{\Theta,\mathcal{E}}(\phi \to 3_1^+)$ (S.E.)	G.O.F. Test Statistic	df	p
0.1	$0.037 - 0.134n^{-0.405}$	0.037 (0.021)	20.7	1	< 0.001
0.2	$0.026 - 0.908n^{-1.065}$	0.026 (0.002)	4.35	1	0.037
0.56	$0.096 - 0.208n^{-0.239}$	0.096 (0.033)	4.65	1	0.031
0.8	$0.098 - 0.257n^{-0.291}$	0.098 (0.004)	0.078	1	0.78
1	$0.136 - 0.264n^{-0.202}$	0.136 (0.020)	0.56	1	0.45
1.5	$0.315 - 0.413n^{-0.083}$	0.315 (0.17)	4.77	1	0.03
2.2	$0.171 - 0.302n^{-0.173}$	0.171 (0.003)	0.01	1	0.94
3.16	$0.262 - 0.370n^{-0.105}$	0.262 (0.034)	0.11	1	0.74
6	$0.281 - 0.173n^{-0.385}$	0.281 (0.173)	7.22	1	0.007
10	$0.185 - 0.315n^{-0.161}$	0.185 (0.071)	2.56	1	0.11

Table 8.13: The results of the fits for the limiting knot transition probabilities of going from $\phi \to 3_1^+$, estimates for this limiting probability with standard error, and statistics pertaining to a goodness of fit test on each regression fit.

consider the cases where the standard error of this estimate is low, namely for $\zeta = 0.2, 0.8, 2$, Figure 8.11 shows that these fits do not fare well outside the region of reliable data.

One solution to this problem might be to loosen the restrictions on the tolerated error defined by ϵ_* in order to get a larger value of $\hat{N}_{\max}(\phi \to 3_1^+|\zeta, A)$. However, increasing $\hat{N}_{\max}$ in such a way may result in considering data that does not truly reflect the intended equilibrium distribution. Another consequence (as one can observe from the results in Table 8.10) is that the larger $\hat{N}_{\max}$ is, the larger the required batch size to obtain essentially independent data becomes. This is because there is generally more correlation between adjacent datapoints corresponding to larger lengths; the consequence of this is that a larger essentially independent batch size is required. Therefore, increasing $\hat{N}_{\max}(*)$ by increasing the tolerated error may not necessarily increase the number of essentially independent datapoints. However, as a main goal of this thesis is to get good fits for limiting knot transition probabilities, it is worth exploring whether increasing the tolerated error will yield a value of $\hat{N}_{\max}(\phi \to 3_1^+|\zeta, A)$ that will be representative of the asymptotic form of $\rho_n^{\Theta,\mathcal{E}}(\phi \to 3_1^+)$. Thus, the tolerated error was increased by changing c in Equation 6.32 from 0.05 to 0.2. This resulted in new estimates for $\hat{N}_{\max}(\phi \to 3_1^+|\zeta, A)$ as well as the essentially independent batch size for each ζ. In Table 8.14 the estimates for $\hat{N}_{\max}(\phi \to 3_1^+|\zeta, A)$ and the essentially independent batch sizes for each ζ are compared for $c = 0.05$ and $c = 0.2$.

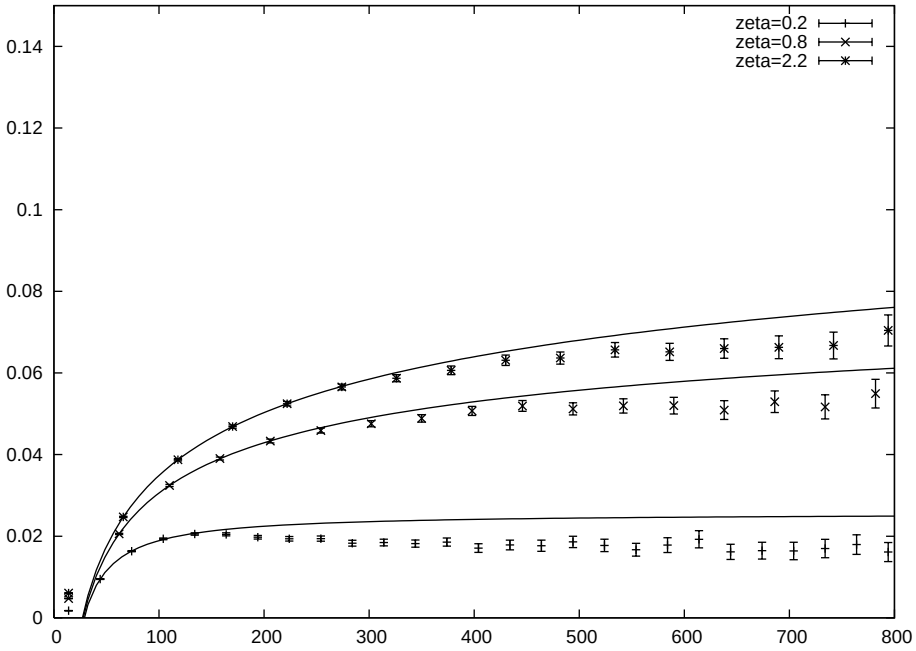

Figure 8.11: An example showing how the 'good' fits of $\rho_n^{\Theta,\mathcal{E}}(\phi \to 3_1^+)$ corresponding to the datapoints in the region of reliable data for $\zeta = 0.2, 0.8, 2.2$ are not accurate for larger values of n.

ζ	$\hat{N}_{\max}(\phi \to 3_1^+ \vert \zeta, A)$ $(c = 0.05)$	Batch Size	$\hat{N}_{\max}(\phi \to 3_1^+ \vert \zeta, A)$ $(c = 0.2)$	Batch Size
0.1	94	20	178	30
0.2	142	30	258	46
0.56	180	38	406	78
0.8	234	48	496	96
1	232	48	608	116
1.5	210	44	508	102
2.2	246	52	604	118
3.16	234	50	546	110
6	240	50	472	96
10	224	48	470	96

Table 8.14: Estimates for the amount of reliable data and the independent batch size corresponding to the limiting knot transition probability $\hat{\rho}_n^{\Theta,\mathcal{E}}(\phi \to 3_1^+)$ when $c = 0.05$ and $c = 0.2$.

One can observe from Table 8.14 that the region of reliable data defined by $\hat{N}_{\max}(\phi \to 3_1^+|\zeta, A)$ approximately doubles when c is increased from 0.05 to 0.2. However, a tradeoff of this is that the essentially independent batch size approximately doubles as well. For $\zeta = 0.1, 0.2, 0.56, 0.8, 1, 2.2$, this increase in c yielded one more essentially independent datapoint. For $\zeta = 1.5, 3.16, 6, 10$ the number of essentially independent datapoints did not change; however, it is expected that a better fit will be achieved because the region of polygon lengths being considered will be more reflective of the asymptotic form of $\hat{\rho}_n^{\Theta,\mathcal{E}}(\phi \to 3_1^+)$. The fits for $\hat{\rho}_n^{\Theta,\mathcal{E}}(\phi \to 3_1^+)$ corresponding to this larger region of reliable data is presented for all ζ values in Table 8.15 and displayed for the ζ values 0.2, 0.8, and 2.2 in Figure 8.12.

ζ	Fit	$\hat{\rho}^{\Theta,\mathcal{E}}(\phi \to 3_1^+)$ (S.E.)	G.O.F. Test Statistic	df	p
0.1	$0.014 + 0.426n^{-1} - 28.54n^{-2}$	0.014 (0.002)	22.6	2	< 0.0001
0.2	$0.016 + 0.908n^{-1} - 64.86n^{-2}$	0.016 (0.001)	14.4	2	0.0007
0.56	$0.045 - 1.986n^{-1.052}$	0.045 (0.001)	3.73	2	0.15
0.8	$0.060 - 0.822n^{-0.740}$	0.060 (0.004)	2.62	2	0.27
1	$0.063 - 1.674n^{-0.891}$	0.063 (0.002)	1.31	2	0.52
1.5	$0.074 - 0.923n^{-0.696}$	0.074 (0.004)	0.57	1	0.45
2.2	$0.078 - 0.865n^{-0.665}$	0.078 (0.002)	0.22	2	0.90
3.16	$0.083 - 0.745n^{-0.610}$	0.083 (0.005)	0.46	1	0.50
6	$0.100 - 0.397n^{-0.401}$	0.100 (0.001)	0.003	1	0.96
10	$0.124 - 0.298n^{-0.270}$	0.124 (0.06)	4.09	1	0.04

Table 8.15: The results of the fits for $\hat{\rho}_n^{\Theta,\mathcal{E}}(\phi \to 3_1^+)$ using a larger region of reliable data, and statistics pertaining to a goodness of fit test on each regression fit.

The ζ values used in Figure 8.12 were chosen to be the same as those used in the fits in Figure 8.11. One can see from Figure 8.12 that the fits of $\hat{\rho}_n^{\Theta,\mathcal{E}}(\phi \to 3_1^+)$ based on the tolerance $c = 0.2$ fare much better for larger values of n than the fits of $\hat{\rho}_n^{\Theta,\mathcal{E}}(\phi \to 3_1^+)$ displayed in Figure 8.11 that are based on the tolerance $c = 0.05$. However, the fits for $\zeta = 0.1$ and 0.2 still fail the goodness of fit test at 5% significance (refer to Table 8.15). It is likely that $\hat{N}_{\max}(\phi \to 3_1^+|\zeta, A)$ is still not large enough in these cases. Although the p-value corresponding to the fit for $\zeta = 0.2$ is small (0.0007), one can observe from Figure 8.12 that this fit fares reasonably well in predicting $\hat{\rho}_n^{\Theta,\mathcal{E}}(\phi \to 3_1^+)$ for larger values of n. With the exception of the fit for $\zeta = 10$, all of the other fits for ζ values larger than 0.2 fare quite well. Figure 8.13 shows the new fits of $\hat{\rho}_n^{\Theta,\mathcal{E}}(\phi \to 3_1^+)$ for each ζ value.

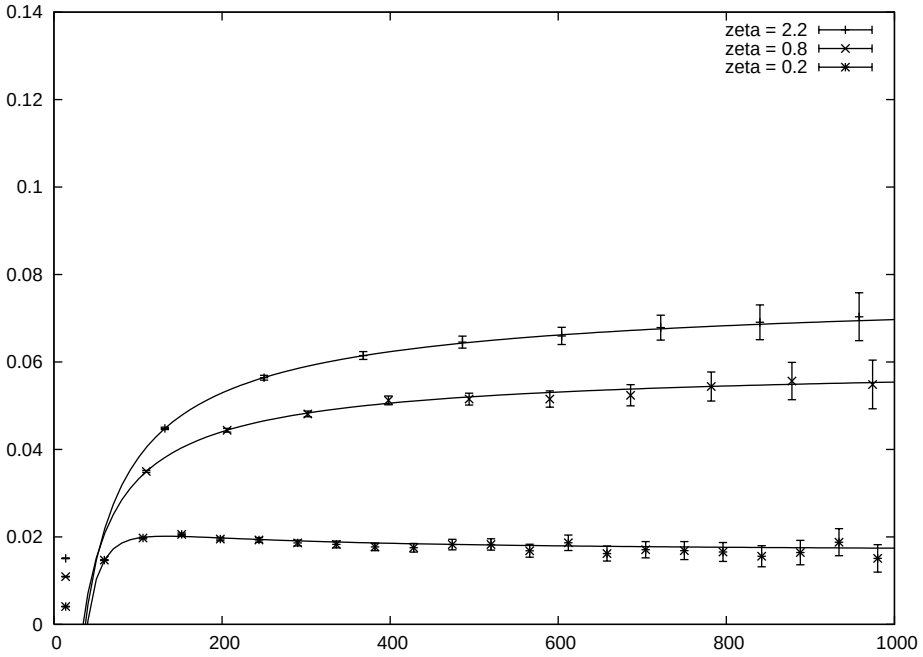

Figure 8.12: Fits of $\rho_n^{\Theta,\mathcal{E}}(\phi \to 3_1^+)$ for $\zeta = 0.2, 0.8, 2.2$ when $\hat{N}_{\max}(\phi \to 3_1^+|\zeta, A)$ is increased.

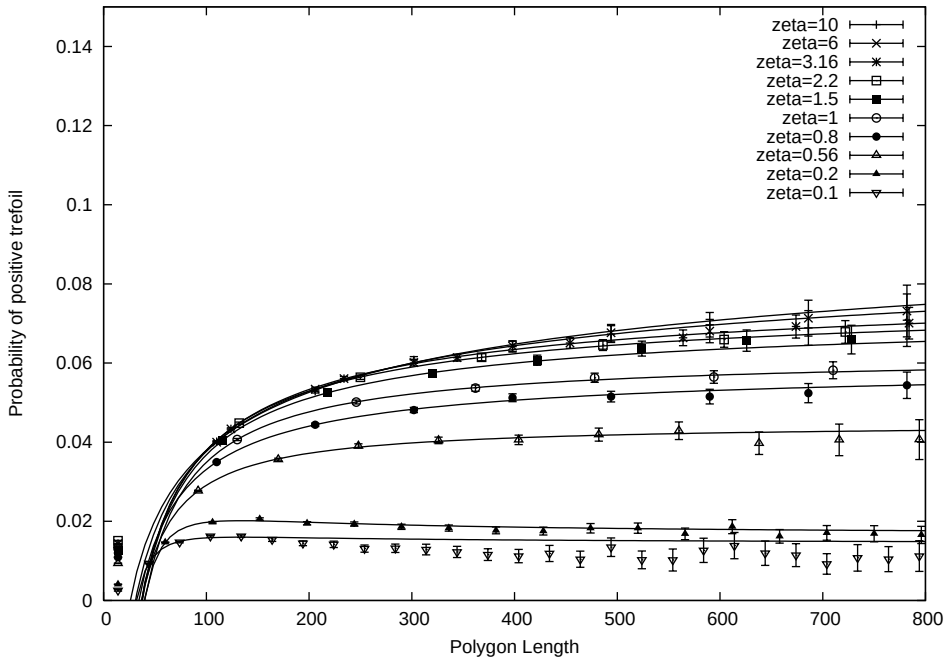

Figure 8.13: Fits of $\rho_n^{\Theta,\mathcal{E}}(\phi \to 3_1^+)$ for all ζ when $\hat{N}_{\max}(\phi \to 3_1^+|\zeta, A)$ is increased.

Recall that the limiting probability of a Θ-SAP staying as an unknot after a successful strand passage decreases as ζ increases. As the trefoil is the most prevalent non-trivial knot that is obtained after a successful strand passage (refer to Table 8.11), we expect that the limiting probability of a Θ-SAP becoming a trefoil after a successful strand passage (*i.e.* $\rho^{\Theta,\mathcal{E}}(\phi \to 3_1^+)$) will increase as ζ increases. The graph shown in Figure 8.13 indicates that $\hat{\rho}^{\Theta,\mathcal{E}}(\phi \to 3_1^+)$ increases with ζ.

8.9 Chapter Summary

The results presented here are for the new I-Θ-BFACF algorithm. The code for the Θ-BFACF algorithm was obtained from Szafron [60, 61] and was modified to include Metropolis sampling based on solvent conditions. Several simulations of the I-Θ-BFACF algorithm were run over a variety of salt concentrations; for each salt concentration 10 independent replications were run for 40 billion time steps each. Each replication is started in a different state with a different random number seed. The sample space for each simulation is the set of all unknotted class II Θ-SAPs.

The average equilibrium polygon length was calculated for each chain and each salt concentration and was displayed in Tables 8.5 and 8.6. From this data, preliminary estimates for the critical values $z_c^{\Theta,\mathcal{E}}(\phi)$ were obtained.

The mean square radius of gyration for each polygon length observed was also calculated for $\zeta = 0.1$ and $\zeta = 10$; Figure 8.1 clearly shows that the mean square radius of gyration grows at a much faster rate for $\zeta = 0.1$ than $\zeta = 10$. The resulting estimates of the mean square radius of gyration for $n = 200$, 300 and 400 from these simulations were compared to the estimates obtained from the I-pivot algorithm simulations. It was found that these estimates from the Θ-BFACF algorithm with energy were slightly lower than the corresponding estimates from the pivot algorithm with energy, but this is expected as the Θ-structure represents two strands of a SAP being pinched together.

Fits for limiting knot transition probabilities were obtained. In some cases such a fit was not useful because there was not enough informative data to get a good fit. In the cases where a successful fit was achieved, it was noticed in Figure 8.9 that as the salt concentration increases, so too does the probability of a knot forming after a successful strand passage. Table 8.11 shows that the majority of these after strand-passage knots are trefoils, and Figure 8.13 indicates that the probability of observing a trefoil after a successful strand passage increases as the salt concentration increases.

Chapter 9

Conclusions/Future Work

9.1 Review

In this work, a new algorithm called the Interacting Θ-BFACF algorithm (or I-Θ BFACF algorithm for short) was developed for sampling random SAPs of varying lengths with a fixed knot type and a fixed structure based on varying solvent conditions. The energy model used to represent these varying solvent conditions contains an attractive force that reflects the solvent quality as well as a screened Coulomb potential that reflects interactions due to the salt concentration of the solution. This energy model was first used for a SAP model by Tesi *et al.* in [64] to study SAPs with a fixed length and variable knot type.

The motivation to have a model reflective of solvent conditions is due to the fact that DNA is negatively charged and interacts with the salt solution in which it exists. A consequence of this interaction is shown by Shaw and Wang in [53] and by Rybenkov *et al.* in [50] where it was observed that the probability of a randomly cyclized DNA molecule being knotted increases with the concentration of salt in the solution.

An independent implementation of the Interacting Pivot Algorithm (or I-Pivot Algorithm for short) based on the energy model as described in [64] was performed, where the trends that were observed with respect to salt concentration were qualitiatively comparable to the simulation results obtained in [64] and the experimental results obtained in [50] and [53].

After these consistency checks, the aforementioned energy model was introduced into the *Local Strand Passage* (*LSP*) model [60, 61, 62, 63]. The LSP model, which forces all SAPs in the model to contain a fixed structure Θ (such SAPs are called Θ-SAPs), represents two strands of a SAP being brought close together for the purpose of a strand passage. Such a strand passage is performed on a Θ-SAP by replacing the Θ-structure with an alternate structure η. The purpose of the LSP model is to study the type II topoisomerase enzyme, which unknots DNA efficiently by performing strand passages on DNA. Because the LSP model is designed to study an enzyme that acts on DNA, and

DNA interacts strongly with the solution in which it exists, it is useful to incorporate the effect of solvent conditions into the LSP model.

In order to sample SAPs from the LSP model with different solvent conditions, it was described how the Θ-BFACF algorithm was modified to incorporate Metropolis sampling based on the energy of a Θ-SAP that is reflective of the chosen solvent conditions. Given a *fugacity* z, a positive integer q and solvent conditions specified by $\mathcal{E}$, the equilibrium probability of obtaining a Θ-SAP ω in this modified algorithm is given by

$$\pi_{\omega}(q, z, \mathcal{E}) := \frac{e^{\frac{-U_{\mathcal{E}}(\omega)}{k_B T}}(|\omega| - 6)|\omega|^{q-1} z^{|\omega|}}{Q^{\Theta}_{K,\mathcal{E}}(z, w)}. \tag{9.1}$$

Code (written in C) for the Θ-BFACF algorithm was provided by the author of [60, 61]. This code was modified to incorporate Metropolis sampling based on SAP energy and resulted in the I-Θ-BFACF Algorithm. Several implementations of the I-Θ-BFACF algorithm were performed using a wide range of salt concentrations. Results for the average polygon length of a chain and the asymptotic behaviour of the probability of successful strand passage and knot transition probabilities were presented. A rough estimate for the radius of convergence of $Q^{\Theta}_{K,\mathcal{E}}(z, w)$ was also presented for each simulation.

9.2 Conclusions

The first goal of this thesis (described by Problem 1) was to find a good way to model ring polymers (in particular circular DNA) in a salt solution. The results presented here for knotting probabilities generated using the I-pivot algorithm (refer to Figure 7.6) show trends that are qualitatively comparable to those obtained in the experiments of Shaw and Wang in [53]; this verifies the observations of Tesi *et al.* in [64]. Although there are still some questions relating to direct comparisons of the SAP model with DNA (described in Future Work), the similarities of the trends observed in simulations compared to those obtained in experimental data suggests that the energy model being used is an effective way to model DNA in solution.

The second goal of this thesis (described by Problem 2) was to study how probabilities relating to strand passages at a random location within a ring polymer change with the concentration of salt in the solution. Assuming that the energy model being used is a good way to model ring polymers in a salt solution, then the strand passage results obtained from simulations of the I-Θ-BFACF algorithm can provide insight into this problem.

The results obtained from the simulations of the I-Θ-BFACF algorithm presented here indicate that as the salt concentration increases, the probability of a successful strand passage decreases (refer to Figure 8.8). This can be explained by the fact that SAPs at higher salt concentrations tend to be more compact, thus causing the vertices around the Θ-structure that must be unoccupied for a successful strand passage to occur to be occupied more frequently. The results from these simulations also indicate that the limiting probabilities of an unknotted Θ-SAP remaining as an unknot after a successful strand passage decreases as the salt concentration increases (see Figure 8.9). Figure 8.10 indicates that the majority of these knotted after-strand passage SAPs are trefoils, and that the probability of obtaining a trefoil after a successful strand passage on an unknotted Θ-SAP increases with the concentration of salt in the solution. These results are reasonable, as one expects the probability of a knot resulting from a random strand passage on an unknotted ring polymer to increase when the polymer is more compact. At high salt concentrations, it was noted in Section 8.8.1 that the limiting probability of obtaining a knot after a successful strand passage is approximately four times larger than in the model that assumes a good solvent.

9.3 Future Work

There are still many questions relating to simulating random Θ-SAPs using the I-Θ-BFACF algorithm based on the energy function described in this work. It would be useful to know exactly how bad the autocorrelation becomes as the average polygon length of a chain increases for small values of ζ (particularly $\zeta = 0.1$). At this point in time it is not known if the results for chain 10 of the simulation where $\zeta = 0.1$ even corresponds to a convergent chain. Will increasing q and removing chain 10 solve this autocorrelation problem?

Because the simulations corresponding to most of the ζ values did not have these autocorrelation problems at the lengths being considered, it would be desirable to generate chains with larger average lengths in order to more accurately estimate the critical z-values $z_c^{\Theta,\mathcal{E}}(\phi)$ for these ζ values. It would also be useful to determine how $z_c^{\Theta,\mathcal{E}}(\phi)$ changes as a function of ζ (for fixed A and v).

In order to obtain a better fit for the limiting knot transition probability of going from an unknot to a trefoil, either more data needs to be generated or larger polygon lengths need to be observed. This can be achieved by running the simulation for longer, and/or by using larger z-values such that the resulting chain remains convergent. Getting a larger region of reliable data is also expected to improve the fits for the limiting successful strand passage probability and the limiting probability

of staying as the unknot.

The ultimate goal of this energy model is to be able to make better comparisons with DNA in solution. One main question that needs to be addressed is "how many base pairs of DNA" correspond to one edge in a SAP? Because SAPs in the lattice have excluded volume, one could compare this to the effective helical diameter of DNA in order to approximate the number of base pairs per edge (as done for one particular salt concentration in Section 7.2.4). However, the effective helical diameter of DNA changes depending on the salt concentration being used; thus, SAPs of a particular length could correspond to a different amount of base pairs of DNA depending on the salt concentration. In order to compare more directly with experimental results, it would be incredibly useful to be able to determine some relationship where a SAP of length a under solvent conditions b represents a DNA chain with c base pairs.

A more immediate goal is to study how the local geometry of the strand passage site impacts these knotting probabilities as a function of salt. This area has been studied by Szafron and Soteros [62, 63] in the good solvent case where they found that there are particular local juxtapositions around the Θ-structure which are particularly favourable or unfavourable to forming a knot after a strand passage (starting with an unknot). As research suggests that type II topoisomerase enzymes do take the local geometry of the strand passage site into account, this seems like a natural path to follow. A goal of future research is to determine solvent and local geometry conditions that will result in a knot reduction factor that is comparable to the 80-fold reduction of knotting seen in the research of Rybenkov *et al.* in [51]. If this goal is achieved, then one can obtain further insight into the mechanism of type II topoisomerases.

References

[1] J. W. Alexander. Topological invariants of knots and links. *Trans. Amer. Math. Soc.*, 30(2):275–306, 1928.

[2] Y. Altun. On the homfly polynomial. *International Mathematical Forum*, 2(56):2753–2757, 2007.

[3] A. D. Bates and A. Maxwell. *DNA Topology.* Oxford University Press, New York, second edition, 2005.

[4] B. Berg and D. Foester. Random paths and random surfaces on a digital computer. *Phys. Lett. B.*, 106:323–326, 1981.

[5] G. Burde and H. Zieschang. *Knots.* de Gruyter. Berlin, 1985.

[6] Y. Burnier, C. Weber, A. Flammini, and A Stasiak. Local selection rules that can determine specific pathways of dna unknotting by type ii DNA topoisomerases. *Nucl. Acids Res.*, 35:5223–5231, 2007.

[7] K. L. Chung. *Markov Chains with Stationary Transition Probabilities.* Springer-Verlag, 1967.

[8] N. Clisby. Efficient implementation of the pivot algorithm for self-avoiding walks. *J. Stat. Phys*, 140:349–392, 2010.

[9] N. Clisby, R. Liang, and G. Slade. Self-avoiding walk enumeration via the lace expansion. *J. Phys. A: Math. Theory.*, 40:10973–11017, 2007.

[10] P. Cromwell. *Knots and Links.* Cambridge University Press, 2004.

[11] C.A. de Carvalho and S. Caracciolo. A new monte carlo approach to the critical properties of self-avoiding random walks. *Phys. Rev. B.*, 27:1635–1645, 1983.

[12] C.A. de Carvalho, S. Caracciolo, and J Frohlich. Polymers and g—s— theory in four dimensions. *Nucl. Phys. B.*, 215:209–248, 1983.

[13] R. Dulbecco and M. Vogt. Evidence for a ring structure of polynoma virus DNA. *Proc. Natl. Acad. Sci. USA*, 50:236–243, 1963.

[14] E. Wasserman E.J. Janse van Rensburg, D. A. W. Sumners and S. G. Whittington. Entanglement complexity of self-avoiding walks. *J. Phys. A. Math. Gen.*, 25:6557–6566, 1992.

[15] P. Freyd et al. A new polynomial invariant of knots and links. *Bull. Amer. Math. Soc.*, 12(2):239–246, 1985.

[16] G. S. Fishman. *Monte Carlo: Concepts, Algorithms and Applications.* Springer-Verlag, 1996.

[17] A. Flammini, A. Maritan, and A Stasiak. Simulations of action of DNA topoisomerases to investigate boundaries and shapes of spaces of knots. *Biophysical Journal*, 87:2968–2975, 2004.

[18] A. Gelman. Inference and monitoring convergence. In Gilks, Richardson, and Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, pages 131–161. Chapman and Hall, 1996.

[19] A. Gelman and D. Rubin. Inference from iterative simulation using multiple sequences. *Stat. Science*, 7:457–472, 1992.

[20] C. J. Geyer. Practical markov chain monte carlo. *Stat. Science*, 7:473–511, 1992.

[21] S. Glasstone. *Introduction to Electrochemistry*. Van Nostrand, Princeton, N.J., 1942.

[22] G. R. Grimmett and D. R. Stirzaker. *Probability and Random Processes*. Oxford Science Publications, 1988.

[23] J. M. Hammersley. On the rate of convergence to the connective constant of the hypercubical lattice. *Quart. J. Math. Oxford*, 12(2):250–256, 1961.

[24] J. M. Hammersley and K. W. Morton. Poor man's monte carlo. *J. Roy. Stat. Soc. B*, 16:23–39, 1954.

[25] Ulrich H. E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Letters*, 281:140–150, 1997.

[26] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57:92–109, 1970.

[27] X. Hua, D. Nguyen, B. Raghavan, J. Arsuaga, and M. Vazquez. Random state transitions of knots: a first step towards modeling unknotting by type ii topoisomerases. *Topology and its Applications*, 154:1381–1397, 2007.

[28] K. Hukushima and K. Nemoto. Exchange monte carlo method and application to spin glass simulations. *J. Phys. Soc. Japan*, 42:281–300, 1996.

[29] V. Jones. A polynomial invariant for knots via von neumann algebra. *Bull. Amer. Math. Noc. (N.S.)*, 12:103–111, 1985.

[30] V. F. R. Jones. Hecke algebra representations of braid groups and link polynomials. *Annals of Math.*, 126:335–388, 1987.

[31] S. Karlin and H. M. Taylor. *A First Course in Stochastic Processes*. Academic Press, 1975.

[32] T. Kennedy. A faster implementation of the pivot algorithm for self-avoiding walks. *J. Stat. Phys*, 106:407–429, 2002.

[33] V. Kocherbitov. Debye screening length. http://www.surfchem.info/calculate/Debye/.

[34] M. Lal. Monte carlo simulations of chain molecules. *Mol. Phys*, 17:64, 1969.

[35] B. Li, N. Madras, and A. D. Sokal. Critical exponents, hyperscaling and universal amplitude ratios for two- and three-dimensional self-avoiding walks. *J. Stat. Phys.*, 80:661–754, 1995.

[36] Z. Liu and H. S. Chan. Efficient chain moves for monte carlo simulations of a wormlike DNA model: excluded volume, supercoils, site juxtapositions, knots and comparisons with random-flight and lattice models. *J. Chem. Phys.*, 128, 2008.

[37] Z. Liu, J. K. Mann, E. L. Zechiedrich E. L., and H. S. Chan. Topological information embodied in local juxtaposition geometry provides a statistical mechanical basis for unknotting by type-2 DNA topoisomerases. *J. Mol. Biol.*, 361:268–285, 2006.

[38] Z. Liu, L. Zechiedrich, and H. S. Chan. Local site preference rationalizes disentangling by DNA topoisomerases. *Phys. Rev. E*, 81, 2010.

[39] N. Madras, A. Orlitsky, and L.A. Shepp. Monte carlo generation of self-avoiding walks with fixed endpoints and fixed length. *J. Stat. Phys*, 58:159–183, 1990.

[40] N. Madras and G. Slade. *The Self-Avoiding Walk*. Birkhuser. Boston, 1996.

[41] N. Madras and A. D. Sokal. The pivot algorithm: A highly efficient monte carl method for the self-avoiding walk. *J. Stat. Phys*, 50:186, 1988.

[42] J. K. Mann. *DNA Knotting: Occurrences, Consequences, and & Resolution*. PhD thesis, The Florida State University, 2007.

[43] J. K. Mann, R. W. Deibler, D. W. Sumners, and E. L. Zechiedrich. Unknotting by type ii topoisomerases. *Abstr. Papers Presented Am. Math. Soc.*, 25:992–187, 2004.

[44] K. C. Neumann, G. Charvin, D. Bensimon, and V. Croquette. Mechanisms of chiral discrimination by topoisomerase iv. *PNAS*, 106:6896–6891, 2009.

[45] E. Orlandini. Monte carlo study of polymer systems by multiple markov chain method. In S. G. Whittington, editor, *Numerical methods for Polymeric Systems*, pages 33–57. Springer-Verlag, 1998.

[46] E. Orlandini, M. C. Tesi, E. J. Janse van Rensburg, and S. G. Whittington. Asymptotics of knotted lattice polygons. *J. Phys. A: Math. Gen.*, 31(28):5953–5967, 1998.

[47] C.W. Patty. *Foundations of Topology*. Jones and Bartlett Publishers, second edition, 2009.

[48] K. Reidemeister. *Knotentheorie*. Springer. Berlin, 1932.

[49] D. Rolfsen. *Knots and Links*. Publish or Perish, Inc., corrected edition, 1990.

[50] V. V. Rybenkov, N. R. Cozzarelli, and A. V. Vologodskii. Probability of DNA knotting and the effective diameter of the DNA double helix. *Proc. Natl. Acad. Sci. U.S.A.*, 90:5307–5311, 1993.

[51] V. V. Rybenkov, C. Ullsperger, A. V. Vologodskii, and N. R. Cozzarelli. Simplification of DNA topology below equilibrium values by type ii topoisomerases. *Science*, 277:690–693, 1997.

[52] R. Scharein. The knotplot site. http://knotplot.com.

[53] S. Y. Shaw and J. C. Wang. Knotting of a DNA chain during ring closure. *Science*, 260:533–536, 1993.

[54] C. K. Singleton, J. Klysik, S. M. Stirdivant, and R. D. Wells. Left-handed Z-DNA is induced by supercoiling in physiological ionic conditions. *Nature*, 299:312–316, 1982.

[55] A. D. Sokal. *Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms*. Cours de Troisime Cycle de la Physique en Suisse Romande, 1989.

[56] C. Soteros, K. Ishihara, K. Shimokawa, M. Szafron, and M. VVazquez. Signed unknotting number and knot chirality discrimation via strand passage. *Prog. Theor. Phys. Supp.*, 191:78–95, 2011.

[57] C. E. Soteros, D. W. Sumners, and S. G. Whittington. Entanglement complexity of graphs in $\mathbb{Z}^3$. *Math. Proc. Camb. Phil. Soc.*, 111:75–91, 1992.

[58] A. Stoimenow. A note on Vassiliev invariants not contained in the knot polynomials. *C. R. Acad. Bulg. Sci.*, 54:9–14, 2001.

[59] D.W. Sumners and S. G. Whittington. Knots in self-avoiding walks. *J. Phys. A: Math. Gen.*, 21(7):1689–1694, 1988.

[60] M. L. Szafron. Monte carlo simulations of strand passage in unknotted self-avoiding polygons. Master's thesis, University of Saskatchewan, 2000.

[61] M. L. Szafron. *Knotting Statistics After A Local Strand Passage In Unknotted Self-Avoiding Polygons*. PhD thesis, University of Saskatchewan, 2009.

[62] M. L. Szafron and C. E. Soteros. The effect of juxtaposition angle on knot reduction in a lattice polygon model of strand passage. *J. Phys. A: Math. Theor.*, 44(32), 2011.

[63] M. L. Szafron and C. E. Soteros. Knotting probabilities after a local strand passage in unknotted self-avoiding polygons. *J. Phys. A: Math. Theor.*, 44(24), 2011.

[64] M. C. Tesi, E. J. Janse van Rensburg, E. Orlandini, D. W. Sumners, and S. G. Whittington. Knotting and supercoiling in circular DNA: A model incorporating the effect of added salt. *Phys. Rev. E*, 49(1):868–872, 1994.

[65] M. C. Tesi, E. J. Janse van Rensburg, E. Orlandini, and S. G. Whittington. Interacting self-avoiding walks and polygons in three dimensions. *J. Phys. A Math. Gen.*, 29:2451–2463, 1996.

[66] M. C. Tesi, E. J. Janse van Rensburg, E. Orlandini, and S. G. Whittington. Monte carlo study of the interacting self-avoiding walk model in three dimensions. *J. Stat. Phys*, 82:155–181, 1996.

[67] J. M. Berger *et al.* Structure and mechanism of DNA topoisomerase ii. *Nature*, 379:225–232, 1996.

[68] M. Tin. Comparison of some ratio estimators. *J. Amer. Stat. Assoc.*, 60:294–307, 1965.

[69] T. Uemura and M. Yanagida. Mitotic spindle pulls but fails to separate chromosomes in type ii DNA mutants. *EMBO Journal*, 5:1003–1010, 1986.

[70] E.J. Janse van Rensburg and S.G. Whittington. The bfacf algorithm and knotted polygons. *J. Phys. A: Math Gen.*, 24:5553–5567, 1991.

[71] M. Vazquez. Private communication.

[72] A. V. Vologodskii. Circular DNA, 1998. www.biophysics.org/portals/1/pdfs/education/vologodskii.pdf.

[73] R. Weil and J. Vinograd. The cyclic helix and cyclic coil forms of polynoma viral DNA. *Proc. Natl. Acad. Sci. USA*, 50:730–739, 1963.

[74] C. C. Wu, T. K. Li, L. Farh, L. Y. Lin, T. S. Lin, Y. J. Yu, T. J. Yen, C. W. Chiang, and N. L. Chan. Structural basis of type ii topoisomerase inhibition by the anticancer drug etoposide. *Science*, 333:459–462, 2011.

Appendix A

Potential Scale Reduction Graphs

The following graphs are the estimated potential scale reduction estimates for the set of 10 replications corresponding to each value of ζ, along with a line at 1.05 corresponding to the maximum point where the replications have considered to have converged to their equilibrium distribution. The estimate of $\tau_{\exp}(\zeta)$ for each graph corresponds to the point where the estimated potential scale is consistently below the cutoff point.

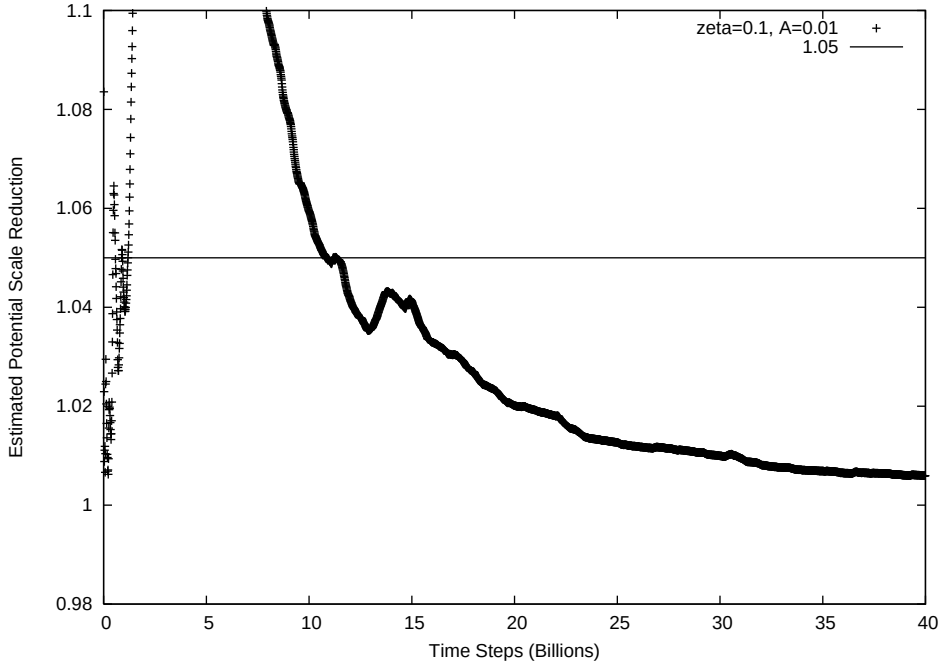

Figure A.1: Estimated potential scale for the simulations corresponding to $\zeta = 0.1$; the estimate $\hat{\tau}_{\exp}(0.1)$ is approximately 12 billion time steps.

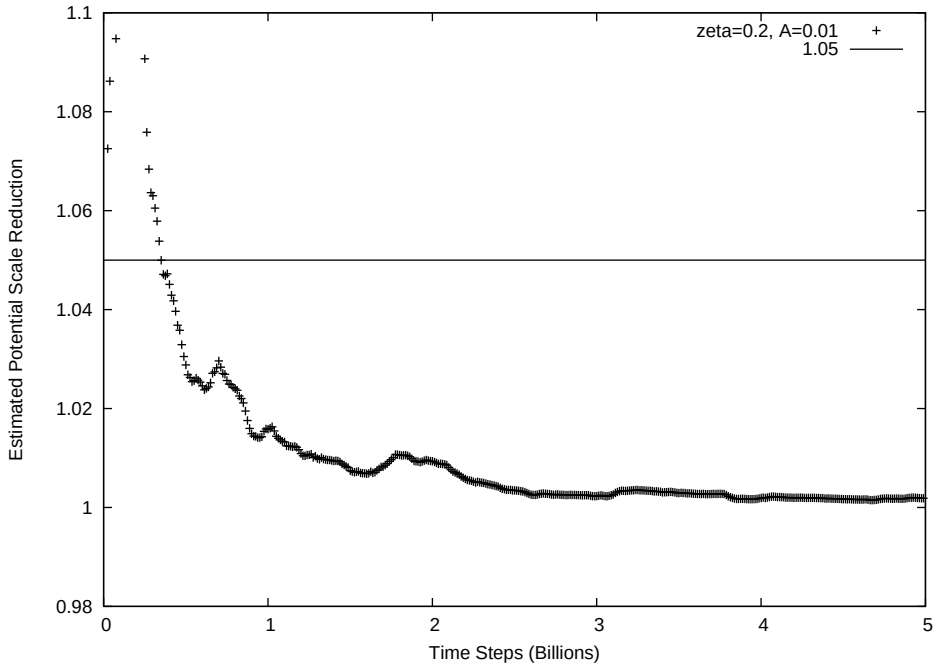

Figure A.2: Estimated potential scale for the simulations corresponding to $\zeta = 0.2$; the estimate $\hat{\tau}_{\exp}(0.2)$ is approximately 0.4 billion time steps.

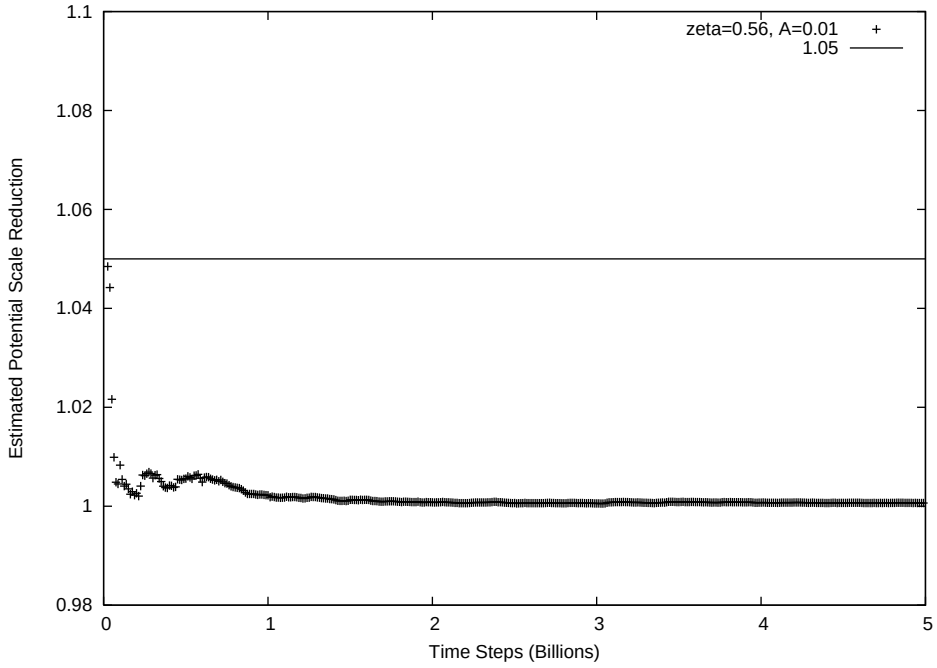

Figure A.3: Estimated potential scale for the simulations corresponding to $\zeta = 0.56$; the estimate $\hat{\tau}_{\exp}(0.56)$ is approximately 0.1 billion time steps.

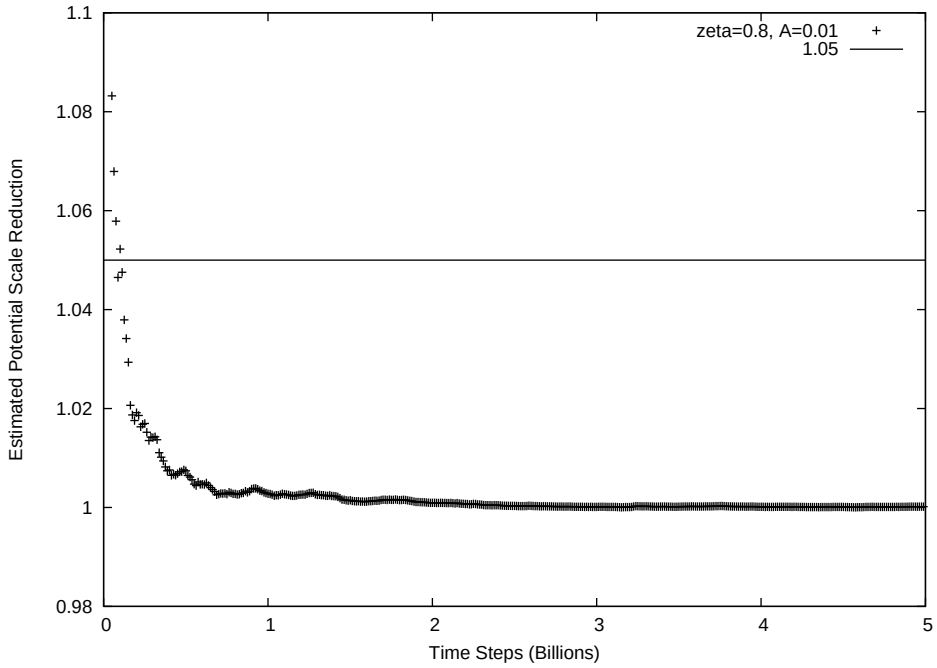

Figure A.4: Estimated potential scale for the simulations corresponding to $\zeta = 0.8$; the estimate $\hat{\tau}_{\exp}(0.8)$ is approximately 0.2 billion time steps.

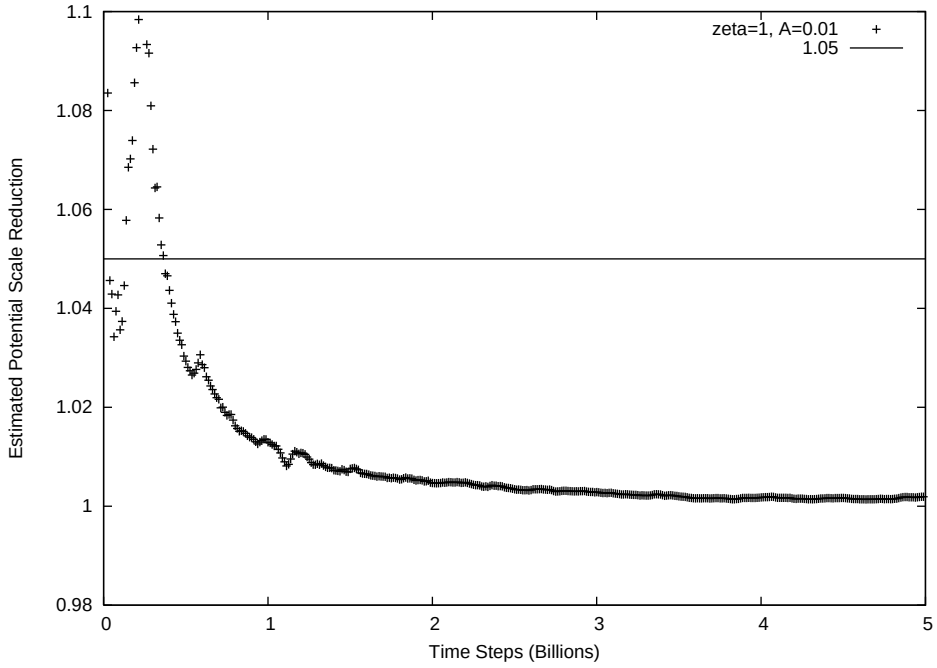

Figure A.5: Estimated potential scale for the simulations corresponding to $\zeta = 1$; the estimate $\hat{\tau}_{\exp}(1)$ is approximately 0.5 billion time steps.

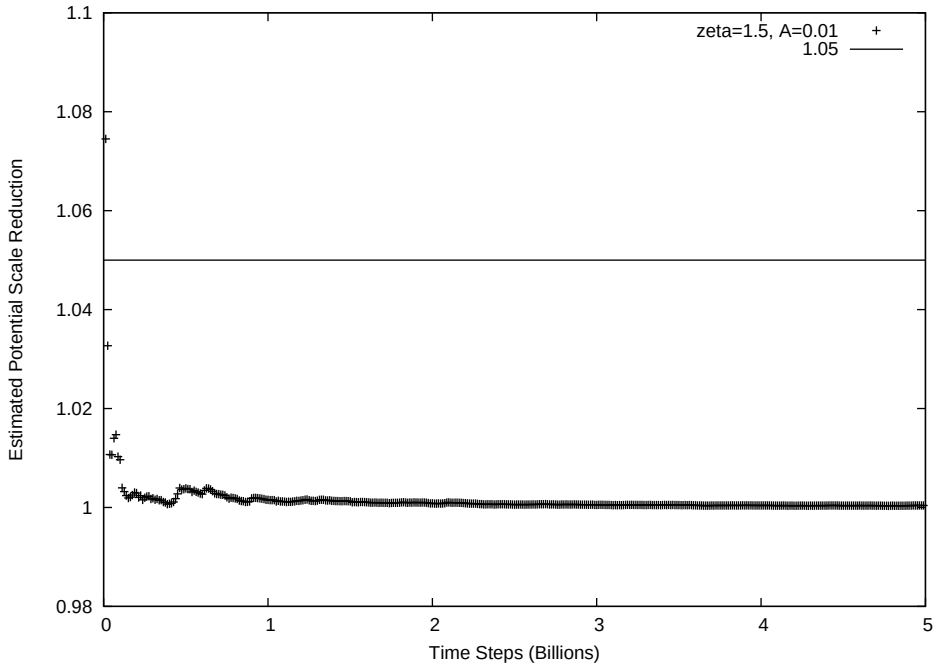

Figure A.6: Estimated potential scale for the simulations corresponding to $\zeta = 1.5$; the estimate $\hat{\tau}_{\exp}(1.5)$ is approximately 0.1 billion time steps.

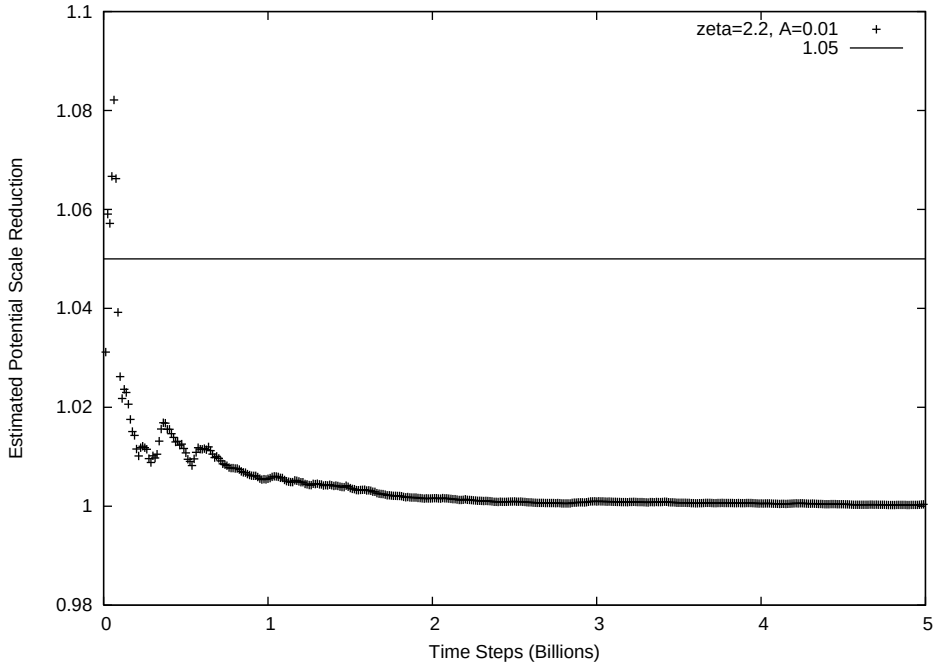

Figure A.7: Estimated potential scale for the simulations corresponding to $\zeta = 2.2$; the estimate for $\hat{\tau}_{\exp}(2.2)$ is approximately 0.2 billion time steps.

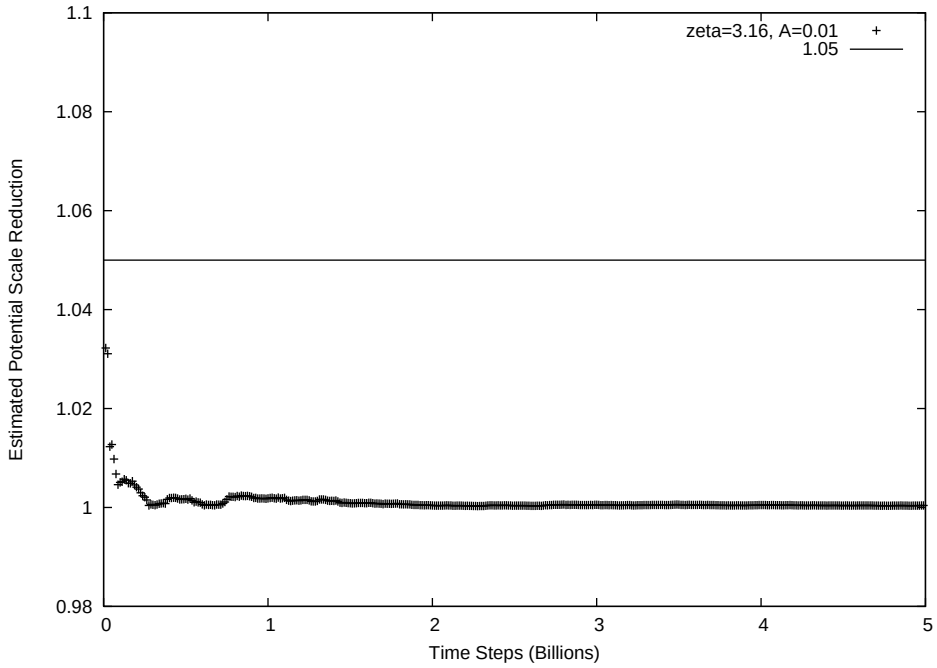

Figure A.8: Estimated potential scale for the simulations corresponding to $\zeta = 3.16$; the estimate $\hat{\tau}_{\exp}(3.16)$ is approximately 0.1 billion time steps.

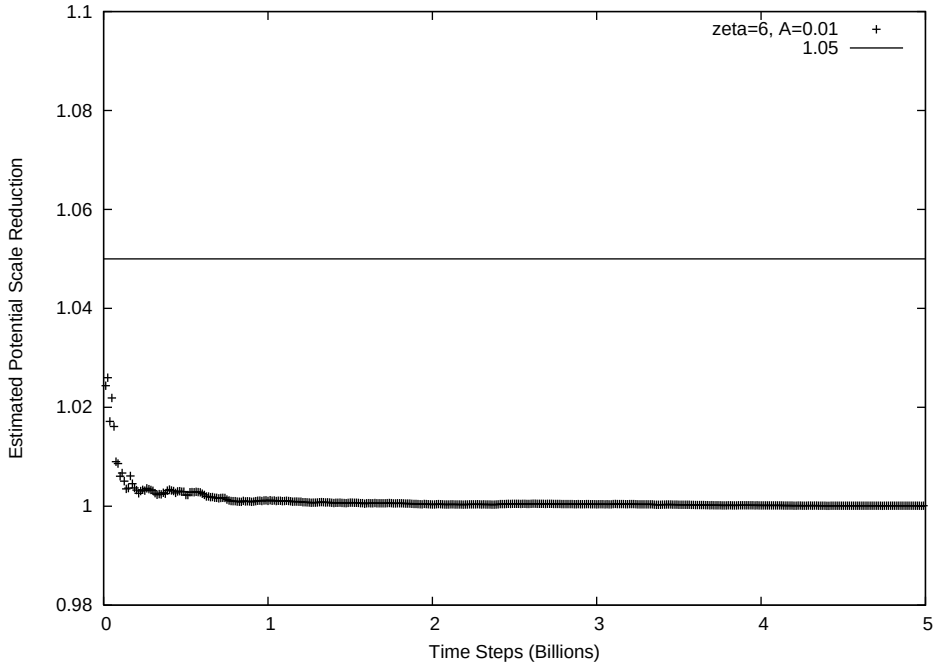

Figure A.9: Estimated potential scale for the simulations corresponding to $\zeta = 6$; the estimate $\hat{\tau}_{\exp}(6)$ is approximately 0.1 billion time steps.

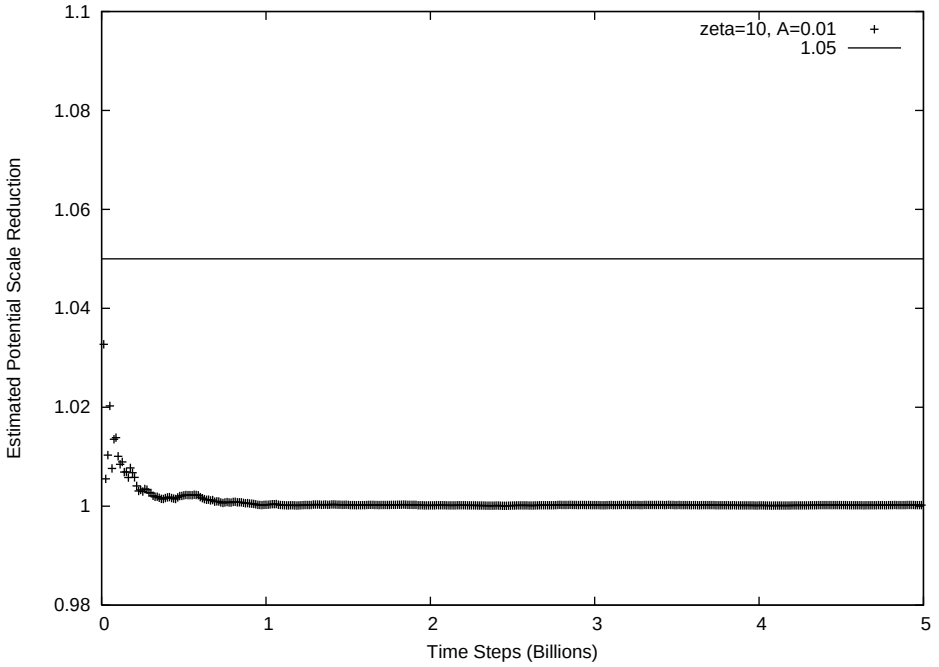

Figure A.10: Estimated potential scale for the simulations corresponding to $\zeta = 10$; the estimate $\hat{\tau}_{\exp}(10)$ is approximately 0.1 billion time steps.

Appendix B

Estimates for $\tau_{\mathrm{INT}}(\zeta, i)$

The following table contains the estimates for the integrated autocorrelation time $\tau_{\mathrm{int}}(\zeta, i)$ for all 10 replications of each value of ζ.

ζ \ Replication (i)	1	2	3	4	5	6	7	8	9	10
0.1	0.35	0.22	0.25	0.72	0.24	0.41	0.80	0.85	0.71	**1.73**
0.2	0.32	0.25	0.14	0.30	0.20	**0.51**	0.23	0.35	0.17	0.30
0.56	< 0.1	0.12	0.12	**0.17**	0.15	< 0.1	0.11	0.13	0.11	0.11
0.8	0.13	0.12	0.09	0.14	0.12	0.08	0.13	**0.18**	0.08	0.07
1	0.09	0.16	0.13	0.19	0.18	0.09	0.09	**0.30**	0.08	0.09
1.5	0.07	0.07	0.10	0.08	0.07	**0.12**	0.07	0.07	0.07	0.06
2.2	0.08	0.08	0.07	0.07	0.08	0.06	0.07	0.08	**0.09**	0.08
3.16	**0.13**	0.07	0.09	0.08	0.06	0.07	0.07	0.09	0.06	0.06
6	0.07	0.08	**0.11**	0.07	0.07	0.06	0.10	0.07	0.06	0.06
10	0.07	< 0.05	0.07	< 0.05	< 0.05	0.06	**0.08**	0.07	0.07	0.07

Table B.1: Estimates for $\tau_{\mathrm{int}}(\zeta, i)$ for each replication and value of ζ; units are in billions of time steps. Bolded entries were the maximum over all 10 replications and were chosen as the estimate for $\tau_{\mathrm{int}}(\zeta)$.

Appendix C

List of Symbols

ϕ - The unknot
3_1 - The trefoil knot
4_1 - The figure-8 knot
3_1^+ - The positive trefoil knot
3_1^- - The negative trefoil knot
Ω_1 - Type I Reidemeister move
Ω_2 - Type II Reidemeister move
Ω_3 - Type III Reidemeister move
$\mathbb{Z}^d$ - The d-dimensional hypercubic lattice
$\mathbb{Z}^3$ - The simple cubic lattice
$|\omega|$ - The number of edges in a SAP ω
$k(\omega)$ - The knot type of ω
$\mathcal{C}_n$ - The set of all n-edge SAWs in $\mathbb{Z}^3$ with $\mathbf{v}_1 = (0,0,0)$
c_n - The number of n-edge SAWs in $\mathbb{Z}^3$ with $\mathbf{v}_1 = (0,0,0)$
$\mathcal{C}$ - The set of all SAWs in $\mathbb{Z}^3$ with $\mathbf{v}_1 = (0,0,0)$
$\mathcal{P}_{2n}$ - The set of all $2n$-edge SAPs in $\mathbb{Z}^3$
p_{2n} - The number of $2n$-edge SAPs in $\mathbb{Z}^3$, up to translation
$\mathcal{P}$ - The set of all SAPs in $\mathbb{Z}^3$
$\mathcal{P}_{2n}(K)$ - The set of all $2n$-edge SAPs in $\mathcal{Z}^3$ with knot type K
p_{2n} - The number of $2n$-edge SAPs in $\mathbb{Z}^3$ with knot type K, up to translation
$\mathcal{P}(K)$ - The set of all SAPs in $\mathbb{Z}^3$ with knot type K
κ - The connective constant for SAWs in $\mathbb{Z}^3$
μ - The growth constant for SAWs in $\mathbb{Z}^3$
κ_ϕ - The connective constant for unknotted SAPs in $\mathbb{Z}^3$
μ_ϕ - The growth constant for unknotted SAPs in $\mathbb{Z}^3$
k_K - The limit inferior defined in Equation 1.8
κ_K - The limit superior defined in Equation 1.8
Θ - A fixed structure for SAPs in $\mathbb{Z}^3$
$V(\Theta)$ - The vertices of Θ
$E(\Theta)$ - The edges of Θ
Θ-SAP - A SAP which contains Θ
σ_1 - Connection class I Θ-SAPs (a.k.a. Θ^+-SAP)
σ_2 - Connection class II Θ-SAPs (a.k.a. Θ^--SAP)
η - The after strand passage structure for Θ-SAPs
$V(\eta)$ - The vertices of η
$E(\eta)$ - The edges of η
$P_{2n}^{\Theta}(K)$ - The set of all class II Θ-SAPs with length $2n$ and knot type K
$p_{2n}^{\Theta}(K)$ - The number of class II Θ-SAPs with length $2n$ and knot type K, up to translation
$P^{\Theta}(K)$ - The set of all class II Θ-SAPs with knot type K
$n(K)$ - The smallest number of edges for which a SAP can have knot type K
$n^{\Theta}(K)$ - The smallest number of edges for which a Θ-SAP can have knot type K
$\rho_{2n}(K)$ - The probability of a random length $2n$ SAP having knot type K

$\rho_{2n}(\bar{\phi})$ - The probability of a random length $2n$ SAP being knotted
$P_{2n}^{\Theta}(s|K)$ - The set of Θ-SAPs with length $2n$ and knot type K for which strand passage is successful
$p_{2n}^{\Theta}(s|K)$ - The number of Θ-SAPs with length $2n$ and knot type K for which strand passage is successful
$P^{\Theta}(K)$ - The set of all class II Θ-SAPs with knot type K
$\rho_{2n}^{\Theta}(s|K)$ - The probability of successful strand passage for a random Θ-SAP with length $2n$ and knot type K
$\mathcal{K}^{\Theta}(K)$ - The set of all knot types that can result from a single strand passage in a Θ-SAP with knot type K
$\mathcal{P}_{2n}^{\Theta}(K'|K,s)$ - The set of all Θ-SAPs in $\mathcal{P}_{2n}^{\Theta}(s|K)$ that have knot type K' after a strand passage
$p_{2n}^{\Theta}(K'|K,s)$ - The number of Θ-SAPs in $\mathcal{P}_{2n}^{\Theta}(s|K)$ that have knot type K' after a strand passage
$\rho_{2n}^{\Theta}(K \to K')$ - The probability of a successful strand passage on a Θ-SAP with knot type K resulting in a SAP with knot type K'
$\rho^{\Theta}(s|K)$ - The limiting successful strand passage probability
$\rho^{\Theta}(K \to K')$ - The limiting knot transition probability of going from knot type K to knot type K' given a successful strand passage
$\kappa_{s|\phi}^{\Theta}$ - The exponential growth rate of $p_{2n}^{\Theta}(s|\phi)$
$\kappa_{K|\phi,s}^{\Theta}$ - The exponential growth rate of $p_{2n}^{\Theta}(K|\phi,s)$
$r^2(\omega)$ - The mean square radius of gyration of ω
$\bar{r}^2(\mathcal{S})$ - The mean square radius of gyration of the elements in $\mathcal{S}$
$C(\omega)$ - The number of contacts in ω
$U_{\mathcal{E}}(\omega)$ - The potential energy of ω with energy parameters $\mathcal{E}$
$r_{ij}(\omega)$ - The euclidean distance between vertices v_i and v_j in ω
ζ - Inverse Debye length, related to salt concentration
v - A negative constant related to solvent quality
$w(n)$ - A polynomial weight function relating to polygon length
z - The fugacity of a chain in the Θ-BFACF algorithm
$Q_K^{\Theta}(z,w)$ - The partition function in the LSP model
$Z_n(\mathcal{E})$ - The partition function in the I-Pivot Algorithm
$Q_{K,\mathcal{E}}^{\Theta}(z,w)$ - The partition function in the ILSP model with energy parameters $\mathcal{E}$
$z_c^{\Theta,\mathcal{E}}(K)$ - The critical z-value in the ILSP model with energy parameters $\mathcal{E}$
$z_c(\mathcal{E})$ - The critical z-value in the ILSP model with energy parameters $\mathcal{E}$
$r(i,j)$ - The probability of swapping chains i and j in a composite Markov chain
$E_\pi(f)$ - The mean of the observable f with respect to π
$\mathrm{var}_\pi(f)$ - The variance of the observable f with respect to π
$\tau_{\mathrm{exp}}(f)$ - The exponential autocorrelation time of the observable f
τ_{exp} - The exponential autocorrelation time of an ergodic Markov chain
$\tau_{\mathrm{exp}}(\zeta)$ - The estimated exponential autocorrelation time for the I-Θ-BFACF simulation with salt concentration ζ
$\sqrt{\hat{R}_j}$ - The estimated potential scale reduction
$\tau_{\mathrm{int}}(f)$ - The integrated autocorrelation time of the observable f
τ_{int} - The integrated autocorrelation time of an ergodic Markov chain
$\tau_{\mathrm{exp}}(\zeta)$ - The estimated integrated autocorrelation time for the I-Θ-BFACF simulation with salt concentration ζ
$E_{z,w,K}[n]$ - The average length of a SAP in a chain of the BFACF algorithm with fugacity z and knot type K
α_{xy} - The Metropolis sampling acceptance probability of going from state x to state y

$\pi_\omega(q, z, \mathcal{E})$ - The equilibrium distribution in the ILSP model
$\hat{\delta}_{2n}^{(r)}(*)$ - The estimated relative standard error of an observable $*$ in replication r
$\hat{\delta}^{(r)}(*)$ - The minimum estimated relative standard error of an observable $*$ in replication r over all polygon lengths n
ϵ_* - The cutoff for choosing $N_{\max}(*)$
$N_{\max}(*)$ - The cutoff for the region of reliable data
$\rho_{2n}^{\Theta,\mathcal{E}}(s|K)$ - The probability of successful strand passage in the ILSP model for a random Θ-SAP with length $2n$ and knot type K
$\rho_{2n}^{\Theta,\mathcal{E}}(K \to K')$ - The probability of a successful strand passage on a Θ-SAP in the ILSP model with knot type K resulting in a SAP with knot type K'
$\rho_{[n_1,n_2]}^{\Theta,\mathcal{E}}(s|K)$ - The grouped-$[n_1, n_2]$ probability of a successful strand passage on a Θ-SAP with knot type K in the ILSP model
$\rho_{[n_1,n_2]}^{\Theta,\mathcal{E}}(K \to K')$ - The grouped-$[n_1, n_2]$ probability of going to knot type K' after a successful strand passage on a Θ-SAP with knot type K in the ILSP model
$\left\langle n_{z_i}(\mathcal{P}^{\Theta}(K)) \right\rangle$ - The estimate for average equilibrium polygon length of a chain of the Θ-BFACF algorithm with fugacity z_i and knot type K
$\bar{n}_{z,\mathcal{E},K}$ - The estimated average equilibrium polygon length for the I-Θ-BFACF simulations with fugacity z, energy parameters $\mathcal{E}$, and knot type K